IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Used Car Price Prediction Using Random Forest Regression

Ms. Punam Khandu Bhaskar[1] and Ms. Shubhangi Pratik Bombale[2]
[1]Student, Jaihind Institute of Management and Research, Junnar, India
[2]Assistatmt Professor, Jaihind Institute of Management and Research, Junnar, India

Abstract : *Estimating the resale value of a pre-owned vehicle is difficult because the price depends on many interacting factors such as brand, age, distance driven, fuel type, transmission, engine capacity and overall condition. Manual appraisal is subjective and often produces inconsistent figures. This paper presents a machine learning approach that predicts used car prices from historical vehicle records using a Random Forest Regressor. The pipeline covers missing-value treatment, categorical encoding, model training and evaluation with Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and the coefficient of determination (R^2). A Flask web application lets a user enter vehicle details, receive an estimated price and explore price trends through charts. Experiments were carried out on a simulated dataset of 4,000 vehicles because a verified public dataset was not used in this study. On a held-out 20% test set the Random Forest achieved an MAE of 0.73 lakh, an RMSE of 2.33 lakh and an R^2 of 0.914, outperforming linear regression and a single decision tree. The system is intended only as a price estimation aid; its output depends on the quality, coverage and relevance of the training data.*

Keywords: used car price prediction, Random Forest, regression, machine learning, Flask, data preprocessing, price estimation

I. INTRODUCTION

The market for pre-owned vehicles has expanded steadily as buyers look for dependable transport at a lower cost than a new purchase. Online listing portals, dealerships and individual sellers all take part in this market, yet a common difficulty remains: deciding a fair asking price. A car's value falls with age and usage, but the rate of that fall differs between brands, body segments and fuel types, and it is further altered by ownership history and physical condition. Because so many variables act together, a simple rule of thumb rarely captures the true market value.

Traditional valuation depends on the experience of a dealer or inspector. Such judgement can be useful, but it varies from person to person and is hard to repeat or audit. A buyer who lacks market knowledge may overpay, while a seller may undervalue a vehicle. A data-driven method that learns from past transactions can offer a consistent reference figure and make valuation more transparent for both parties.

Supervised machine learning suits this task well, because the target (selling price) is a continuous number and past records supply labelled examples. Ensemble tree methods are especially attractive for tabular data, since they model non-linear effects and interactions between features without heavy feature scaling [1], [3]. This work adopts the Random Forest Regressor and wraps it in a lightweight web interface so that the model can be used without any programming knowledge.

The main objectives of the study are: (i) to prepare vehicle data through cleaning and encoding; (ii) to build and tune a Random Forest model for price prediction; (iii) to compare it with a linear baseline and a single decision tree using MAE, RMSE and R^2; (iv) to identify which attributes influence price most; and (v) to deploy the trained model in a Flask application with interactive visualisations.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

II. PROBLEM STATEMENT

Setting a fair resale price for a used car is still largely a matter of personal judgement. Sellers and dealers usually quote a figure based on experience, a few comparable advertisements or a rough rule about yearly depreciation, and buyers often have no reliable way to check it. Because price depends jointly on age, brand, distance driven, fuel type, transmission, engine capacity, condition and ownership history, two people appraising the same car can arrive at noticeably different values. This inconsistency leads to overpricing, undervaluation and long negotiations, and it reduces trust in the market.

The problem addressed in this paper can be stated formally as follows: given a set of historical used-car records, each described by a feature vector x of vehicle attributes and a known selling price y, learn a function f such that f(x) gives an accurate estimate of the selling price of a previously unseen vehicle, with the estimation error measured by MAE, RMSE and R^2. In addition, the learned function must be made available through a simple web interface so that a person with no technical background can obtain an estimate by entering vehicle details.

Three specific difficulties make the task non-trivial: the relationship between price and its drivers is non-linear (for example, depreciation is steep in early years and slower later); real data contain missing and mixed-type values that must be cleaned and encoded; and a single model fitted to a skewed price distribution tends to perform poorly on expensive cars. The proposed system is designed to handle each of these issues.

III. RELATED WORK

Early studies on vehicle price estimation relied largely on regression. Pudaruth [4] compared several algorithms, including multiple linear regression, k-nearest neighbours, naive Bayes and decision trees, on car advertisements collected from a local market, and observed that the accuracy of every method was limited by the small number of records and the diversity of vehicle models. Gegic et al. [5] combined several learners in a mixed approach and reported that ensemble schemes gave better estimates than individual models on their data.

The Random Forest, introduced by Breiman [1], builds many decision trees on bootstrap samples and averages their outputs. Averaging lowers variance and reduces over-fitting compared with one deep tree, while the random choice of candidate features at each split decorrelates the trees. Hastie et al. [3] describe these properties in detail and note that forests perform well with mixed numerical and categorical inputs. Software libraries such as scikit-learn [2] have made such methods easy to apply, and the Flask micro-framework [8] provides a simple way to publish trained models as web services.

Most earlier work stops at reporting model accuracy. The present paper goes one step further by connecting the trained estimator to a browser-based form and a small set of charts, so that the result is directly usable by a non-technical person.

IV. DATASET AND METHODOLOGY

A. Dataset Description

A tabular dataset of 4,000 used-car records was generated for this study using a controlled simulation. Brand shares, vehicle ages, distance driven, fuel type, transmission, engine capacity, condition and ownership count were drawn from plausible distributions, and the selling price was produced from a depreciation relationship with random noise added to imitate real market variation. Because the data are simulated, the numerical results reported here demonstrate the workflow and relative model behaviour; they should not be read as measurements of the real market. Table I lists the attributes.

TABLE I: ATTRIBUTES OF THE DATASET

Attribute	Type	Description	Values / Range
Brand	Categorical	Manufacturer of the car	8 brands (Maruti to Audi)
Year	Numeric	Year of manufacture	2008 – 2024
Km_Driven	Numeric	Distance covered	2,000 – 300,000 km

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Fuel	Categorical	Fuel type	Petrol, Diesel, CNG
Transmission	Categorical	Gearbox type	Manual, Automatic
Engine_cc	Numeric	Engine capacity	800 – 3,000 cc
Condition	Ordinal	Overall state of vehicle	Fair, Good, Excellent
Owners	Numeric	Number of previous owners	1 – 3
Price_Lakh	Numeric (target)	Selling price	1.05 – 68.97 lakh INR

B. Data Preprocessing

Real listings are rarely complete, so missing entries were deliberately introduced at low rates to test the cleaning stage: 83 values of engine capacity, 54 of kilometres driven and 40 of condition were absent (about 4.4% of rows affected in total). Numeric gaps were filled with the column median, which is robust to the few very expensive cars, and the missing condition labels were filled with the most frequent category. The manufacturing year was converted to vehicle age (2025 minus year), because age relates to depreciation more directly. Condition was mapped to an ordered scale (Fair = 0, Good = 1, Excellent = 2) and transmission to a binary flag. Brand and fuel type were one-hot encoded, giving 17 input columns in total. Descriptive statistics of the numeric fields are given in Table II.

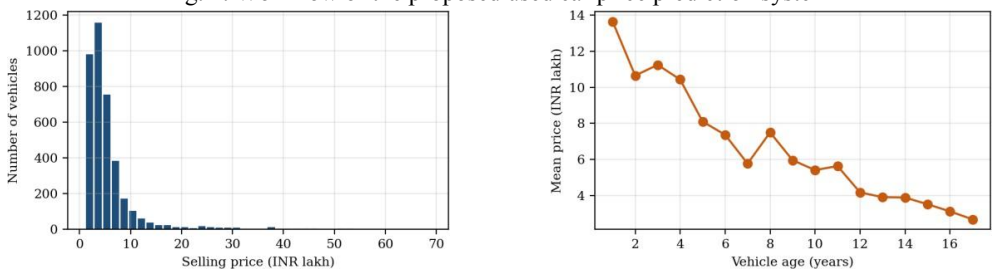
TABLE II: DESCRIPTIVE STATISTICS OF NUMERIC ATTRIBUTES

Attribute	Mean	Std. Dev.	Min	Max
Year	2016.0	4.82	2008	2024
Km_Driven (km)	98,184	66,538	2,000	300,000
Engine_cc (cc)	1,512	478	800	3,000
Owners	1.47	0.66	1	3
Price (lakh INR)	6.63	8.32	1.05	68.97

The price distribution is strongly right-skewed (Fig. 2a): most cars sell below ten lakh, while a small premium segment extends the tail. Average price falls quickly with age (Fig. 2b), and brand tier has a clear effect on price (Fig. 3a). The correlation matrix (Fig. 3b) shows a negative link between price and both age and distance driven, and a positive link with engine capacity.



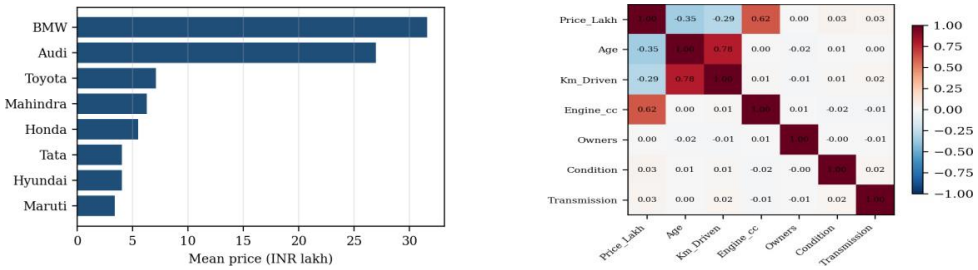
Fig. 1. Workflow of the proposed used car price prediction system



(a) Price distribution (b) Mean price versus vehicle age

Fig. 2. Exploratory analysis of the price variable

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2



(a) Mean price by brand (b) Correlation matrix

Fig. 3. Brand-wise price comparison and feature correlations

C. Random Forest Regression

A Random Forest regressor with B trees is trained on B bootstrap samples drawn from the training set. For each split inside a tree, only a random subset of the features is examined, and the split that most reduces the within-node variance is chosen. The final prediction for a vehicle x is the mean of the predictions of all trees:

$$\hat{y}(x) = (1/B)\,\Sigma\, T_b(x),\ b = 1,\ldots,B \quad (1)$$

In this work B = 200 trees were used, the minimum number of samples in a leaf was set to 2, and the random seed was fixed at 42 for reproducibility. The data were divided into 80% for training (3,200 records) and 20% for testing (800 records). For comparison, a multiple linear regression model and a single decision tree limited to a depth of 12 were trained on the same split.

D. Evaluation Metrics

Model quality was judged with three standard regression measures [6], [7]. For n test vehicles with actual prices y_i, predicted prices $\hat{y}_i$ and mean actual price $\bar{y}$:

$$MAE = (1/n)\,\Sigma\,|y_i - \hat{y}_i| \quad (2)$$

$$RMSE = \sqrt{[\,(1/n)\,\Sigma\,(y_i - \hat{y}_i)^2\,]} \quad (3)$$

$$R^2 = 1 - \Sigma\,(y_i - \hat{y}_i)^2 / \Sigma\,(y_i - \bar{y})^2 \quad (4)$$

MAE gives the average size of the error in the same unit as the price, RMSE penalises large mistakes more strongly, and R^2 states the proportion of price variance explained by the model.

V. SYSTEM IMPLEMENTATION

The system was written in Python. The pandas library [9] handled data preparation, scikit-learn [2] supplied the models and metrics, and Matplotlib produced the charts. After training, the fitted model and the list of encoded column names were saved to disk. The Flask application [8] loads these files once at start-up and exposes three pages: a home page, a prediction form and a visualisation dashboard.

On the prediction page the user chooses the brand, enters the year, kilometres driven and engine capacity, and selects fuel type, transmission, condition and number of owners. The server rebuilds a single input row with exactly the same encoding used in training, passes it to the model and returns the estimated price. Input validation rejects impossible values, for example a manufacturing year in the future or a negative distance. The dashboard page presents price distribution, price against age and brand-wise averages so that users can see how an estimate compares with general trends.

VI. RESULTS AND DISCUSSION

A. Model Comparison

Table III reports the test-set performance of the three models. The Random Forest gave the lowest MAE and RMSE and the highest R^2. Its MAE of 0.73 lakh is less than half the 1.59 lakh of linear regression, because depreciation is non-

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

linear and interacts with brand and engine size, relationships a straight-line model cannot express. The single decision tree improved on the linear model in MAE but its RMSE stayed high, which indicates unstable predictions on expensive cars; averaging many trees corrected this.

TABLE III: TEST-SET PERFORMANCE OF THE MODELS

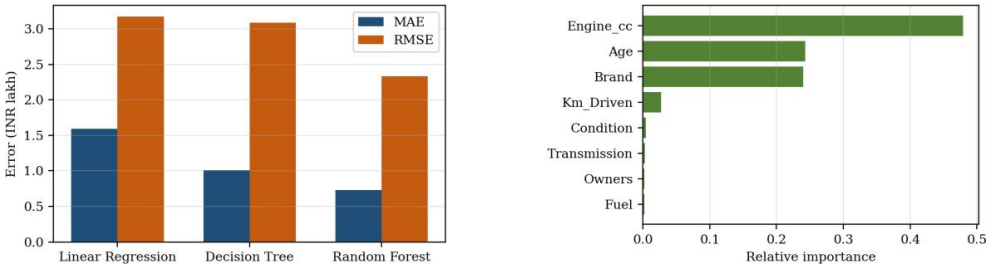
Model	MAE (lakh)	RMSE (lakh)	R^2 score
Linear Regression	1.59	3.17	0.841
Decision Tree (depth 12)	1.01	3.08	0.849
Random Forest (200 trees)	0.73	2.33	0.914

B. Cross-Validation

To check that the result does not depend on one particular split, five-fold cross-validation was performed on the full dataset (Table IV). The R^2 values ranged from 0.943 to 0.976 with a mean of 0.964 and a standard deviation of 0.011. The single-split R^2 in Table III is lower than this range most likely because the random 20% test set contains a number of high-priced premium cars, whose absolute errors are naturally larger. The spread across folds (0.943 to 0.976) also suggests that the premium segment is the least stable part of the model.

TABLE IV FIVE-FOLD CROSS-VALIDATION OF THE RANDOM FOREST

Fold	1	2	3	4	5
R^2 score	0.9756	0.9733	0.9654	0.9433	0.9625



(a) Error comparison of the models (b) Grouped feature importance

Fig. 4. Model error comparison and relative importance of vehicle attributes

C. Feature Importance

Grouped feature importances are shown in Fig. 4b. Engine capacity (47.9%), vehicle age (24.3%) and brand (24.0%) together account for more than 95% of the total importance. Engine capacity ranks first partly because, in the simulated data, it also separates economy cars from premium ones, so it acts as a stand-in for vehicle segment. Distance driven contributes about 2.7%, and fuel type, transmission, condition and ownership each contribute under 0.5%. In real data, condition and ownership may matter more than shown here, which is why the importance ranking should be re-examined after training on genuine records.

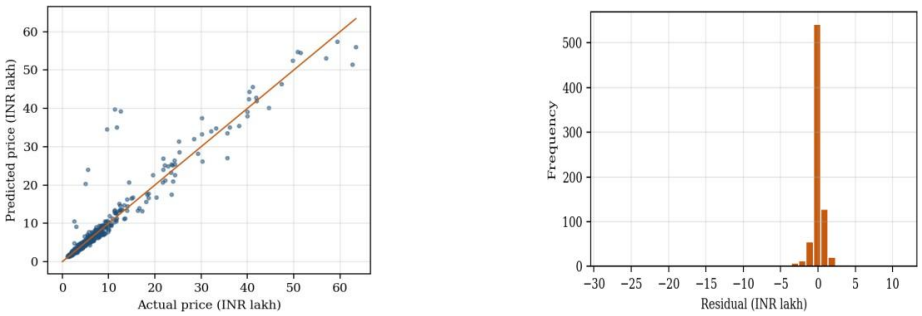
D. Worked Calculation of Metrics

The metrics can be verified by hand on five test vehicles (Table V). The absolute errors are 0.08, 0.15, 4.50, 0.01 and 1.17 lakh, whose sum is 5.91, so MAE = 5.91 / 5 = 1.18 lakh. The squared errors sum to 21.64, giving MSE = 21.64 / 5 = 4.33 and RMSE = $\sqrt{4.33}$ = 2.08 lakh. The mean actual price is 66.45 / 5 = 13.29 lakh and the total variation $\Sigma(y - \bar{y})^2$ is 1,258.21, so $R^2 = 1 - 21.64 / 1{,}258.21 = 0.983$. The large error of 4.50 lakh arises from the Audi, which is the most expensive car in the sample and illustrates why RMSE is higher than MAE.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

TABLE V: SAMPLE PREDICTIONS ON TEST VEHICLES (INR LAKH)

Brand	Year	Km driven	Fuel	Engine (cc)	Actual	Predicted	Abs. error
Maruti	2022	29,700	Diesel	1,400	5.44	5.52	0.08
Maruti	2009	176,800	Diesel	1,000	1.76	1.61	0.15
Audi	2021	54,700	Petrol	3,000	44.61	40.11	4.50
Honda	2016	107,500	Petrol	1,800	4.93	4.92	0.01
Honda	2022	43,700	Petrol	1,800	9.71	8.54	1.17



(a) Actual versus predicted price (b) Residual distribution

Fig. 5. Prediction quality of the Random Forest on the test set

E. Error Analysis

Fig. 5a shows that predictions follow the ideal diagonal closely for cars below twenty lakh, with wider scatter for the premium segment where training examples are few. The residuals in Fig. 5b are centred near zero and roughly symmetric, so the model does not systematically over- or under-estimate. The mean absolute percentage error on the test set is about 10.7%; for budget and mid-range cars, where most transactions occur, the relative error is smaller than for luxury cars.

F. Effect of Hyper-Parameters

The influence of the number of trees was examined by training forests of 10 to 300 trees with all other settings fixed (Table VI). Error falls sharply up to about 100 trees and then changes only slightly, so 200 trees were kept as a balance between accuracy and training time. The minimum leaf size was also varied: sizes of 1, 2, 5 and 10 gave R^2 values of 0.919, 0.914, 0.939 and 0.901 respectively. A leaf size of 5 gave the highest R^2 and the lowest RMSE (1.96 lakh) on this particular split, while a leaf size of 1 gave the lowest MAE (0.72 lakh). These differences are small, and because they come from a single split they should be confirmed by cross-validated tuning before any change is adopted. The main results in this paper therefore use the default-style setting of 200 trees and a leaf size of 2.

TABLE VI: EFFECT OF NUMBER OF TREES (LEAF SIZE = 2)

Trees	MAE (lakh)	RMSE (lakh)	R^2 score
10	0.82	2.65	0.889
25	0.76	2.54	0.898
50	0.74	2.42	0.907
100	0.73	2.33	0.914
200	0.73	2.33	0.914
300	0.73	2.24	0.920

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

VII. LIMITATIONS

Several limitations should be stated clearly. First, the dataset is simulated, so the conclusions about feature importance and absolute error cannot be transferred directly to a real market. Second, the model uses only structured attributes and ignores factors such as accident history, service records, colour, city of registration and seasonal demand. Third, a forest cannot extrapolate beyond the range of its training data, so very new or very rare models will be estimated poorly. Finally, market prices change over time, and a model trained once will become outdated unless it is retrained periodically. The estimate produced by the application is therefore a guide, not a binding valuation.

VIII. CONCLUSION AND FUTURE SCOPE

This paper developed a used car price estimation system based on the Random Forest Regressor and delivered it through a Flask web interface. After median and mode imputation and suitable encoding, the forest explained about 91% of price variance on held-out data and reduced the mean absolute error by more than half relative to linear regression. Vehicle age, brand and engine capacity were the dominant predictors in the experiment. The accompanying dashboard helps users place an estimate in the context of wider price trends.

In future work the system will be trained on verified listings from real marketplaces, hyper-parameters will be tuned using grid or randomised search, and gradient boosting methods will be compared with the forest. Predicting the logarithm of price may reduce the effect of expensive outliers, and adding images or inspection reports could capture condition more objectively. Periodic retraining and prediction intervals would further improve the reliability of the estimates.

REFERENCES

[1] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.
[4] S. Pudaruth, “Predicting the price of used cars using machine learning techniques,” International Journal of Information & Computation Technology, vol. 4, no. 7, pp. 753–764, 2014.
[5] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, “Car price prediction using machine learning techniques,” TEM Journal, vol. 8, no. 1, pp. 113–118, 2019.
[6] C. J. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” Climate Research, vol. 30, no. 1, pp. 79–82, 2005.
[7] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature,” Geoscientific Model Development, vol. 7, no. 3, pp. 1247–1250, 2014.
[8] M. Grinberg, Flask Web Development: Developing Web Applications with Python, 2nd ed. Sebastopol, CA, USA: O’Reilly Media, 2018.
[9] W. McKinney, “Data structures for statistical computing in Python,” in Proc. 9th Python in Science Conf. (SciPy), 2010, pp. 56–61.