Machine Learning-Based Approach for Used Car Price Prediction

Vijaykumar Bhanuse
Vishwakarma Institute of Technology
Pune, India

Prapti These
Vishwakarma Institute of Technology
Pune, India

Tejas Runwal
Vishwakarma Institute of Technology
Pune, India

Rohit Patil
Vishwakarma Institute of Technology
Pune, India

Rutuja Sarode
Vishwakarma Institute of Technology
Pune, India

Nandini Perlawar
Vishwakarma Institute of Technology
Pune, India

Abstract: The previously owned vehicle market around the world has seen very high growth, which demands more accurate and reliable methods to estimate the prices of used cars. Traditional valuation has provided inconsistent results because of its subjectivity. This study presents the use of machine learning (ML) algorithms, XGBoost, CAT Boost, and Extra Trees in predicting the resale value for used vehicles. There are other sections in this study that review existing literature, key features (like mileage, brand, model, fuel type, and transmission) impacting car prices and performance evaluation of selected algorithms using a curated dataset. Experimental results show that ensemble learning models greatly improve prediction accuracy in capturing nonlinear patterns and interactions in the data. The paper ends with future scope, such as using large data, better feature engineering, and explainable AI to develop even more robust and interpretable prediction models.

Keywords- Price prediction, used cars, Machine learning, Model, Resale value, Sellers.

I. Introduction

In recent years, the pre-owned automobile market has witnessed exponential growth, driven by factors such as increasing vehicle prices, improved vehicle longevity, and a growing preference for affordable transportation options. As a result, accurately estimating the resale value of used cars has become a critical component for buyers, sellers, and dealers alike. Traditional methods of pricing rely heavily on subjective assessment and basic depreciation models, often leading to inconsistencies and inaccuracies.

With the advent of machine learning (ML), data-driven approaches have shown significant promise in enhancing prediction accuracy by uncovering hidden patterns and complex relationships between vehicle attributes and their market value. ML algorithms are capable of processing a wide range of numerical and categorical inputs—such as mileage, fuel type, brand, model, and year of registration—to generate robust prediction models. This has paved the way for intelligent decision-making systems in the automotive resale sector.

This paper aims to explore and synthesize existing research efforts in the domain of used car price prediction using supervised machine learning techniques. By reviewing methodologies, algorithms, data types, outcomes from multiple studies, and its future scope this work provides insights into prevailing trends and research gaps in the field.

II. Literature Review

We reviewed a total of twenty research articles about the prediction of used car prices by applying different AI and machine learning techniques. From these, we shortlisted and kept relevant ones that made use of analytical

models such as linear regression, decision trees, random forests, support vector machines, neural networks, or deep learning techniques in an attempt to study the different research studies.

We would have organized the papers based on the method or the datasets they were set against, as well as the metrics they used to validate their performance. Finally, we provided an account of strengths and weaknesses, accuracy outcomes, and other consequences that were available. The purpose was to show trends in the field, to see which models proved to be the highest performer, and eventually to discuss gaps or challenges that still exist.

We also reduced the methodology, model performances, and most significant findings of each paper into a very structured and brief outline that allowed us to follow the evolution of methods used in car price prediction so that we could draft a well-organized and thorough literature review.

Empirical findings from an extensive review of recent works indicate that car price prediction has become a promising research area wherein several supervised learning algorithms have been employed to improve model accuracy and robustness. Among the twenty papers collected, some employed datasets that had key parameters such as year of manufacture in [1] [2], mileage in [1][6], fuel type in [4][5], transmission [5][7], colour[2][3], and model [1][10]. Among these parameters, mileage, brand, and model are the most correlated which affected the resale value prediction. These attributes are a mixture that sees both quantitative and qualitative variables affecting the condition of the particular vehicle involved and the perception of consumers with regards to that vehicle.

With respect to the methodology, Multiple Linear Regression [11][14], Decision Trees Regression [5][8], Random Forests [8][13], Support Vector Machines [18][11], K-Nearest Neighbours (KNN) [2][7], are few of the most innovatively inducted algorithms. Each has its own merit- for instance, regression trees are favoured due to the interpretability and ability to cope with nonlinear interactions; KNN works wonders in patterning mapping within the local neighbourhood of the data. This particularity also holds for the papers under review, with much variance in the selection of data types. Most studies deal with a combination of numerical [12][17] and categorical data [15][16], and hence, the preprocessing methods such as feature encoding and scaling become highly demanding. Given these set challenges, XGBoost and CAT Boost perform superbly despite the absence of any drop in accuracy and performance.

Most studies have common constraints like constrained datasets, less than optimal feature engineering, and inability to generalize geographically their models. Such constraints lock down possible areas of future exploration: growth of dataset, application of cross-validation methods, and honing in on hybrid learning or ensemble methods. The literature reveals that machine learning has the potential of upgrading used car price estimation from a not-so-reliable/less scalable process.

III. Methodology

The study followed a typical machine learning process to predict the prices of used vehicles, consisting of activities such as data collection, data preprocessing, model training, and finally, model evaluation.

1. Approach and Implementation

a) Data Collection:

The used-car data were collected publicly and contained attributes such as brand name, model name, year of manufacture, fuel type, transmission type, kilometres driven, engine size, and selling price.

b) Data Preprocessing:

Missing values were treated accordingly, and categorical variables were encoded using appropriate methods such as label encoding or one-hot encoding. New features such as car age were derived from the year of manufacture. Outliers were then treated to improve the reliability of the model.

c)Feature Selection:

Correlation and relevance to the domain were employed to find the more significant features and lessen noise and variance from the overall data, allowing a more accurate prediction.

d) Model Development:

Regression techniques were implemented, including Linear Regression, Decision Trees, Random Forest, and Extra Trees. The model was fitted on 80% of the data while holding the remaining 20% for testing, setting the course for hyperparameter tuning performed through cross-validation.

e) Model Evaluation:

The assessment parameters R^2 score, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were used, with visual prediction vs actual plot augmenting the numerical assessments.

f) Model Comparison

All models were compared based on performance evaluation metrics, where ensemble methods outdid others in predictive accuracy and generalization ability.

2. Dataset Selection:

The *table 2.1* contains information about price prediction of used cars sales depending upon various features. It aims to help buyers compare used cars based on key factors like price, mileage, fuel type, and condition. It makes it easier to choose a car that fits their budget and needs efficiently.

Features:

Name: Name and model of that particular car.

Location: The city where cars are available.

Year: Year of Manufacture
Kilometres Driven: Distance of car has been driven (in km)

Fuel Type: Type of fuel used (CNG, Diesel, Petrol, LPG)

Transmission: Gear System (Manual or Automatic)

Owner Type: Indicates if the seller is the first or second owner

Mileage: Fuel Efficiency

Engine: Engine Capacity (in CC)

Power: Engine Power output (in bhp)

Seats: Number of seats in the vehicle

New price: The price of car when new (if available)

Target Variable:

Price: Current selling price (in Lakhs)

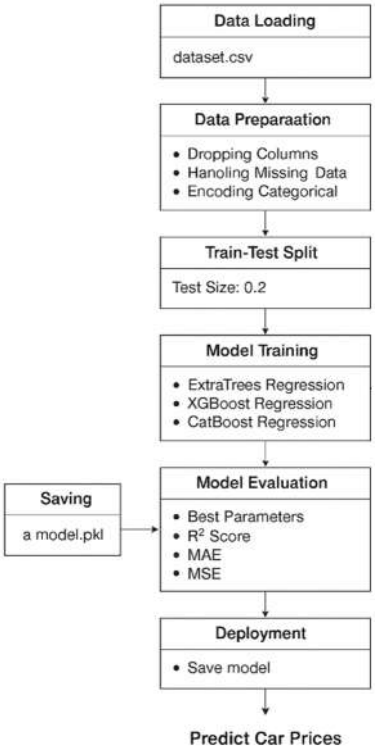

Figure 2.1: Resale car price prediction flowchart

3. Algorithms Used:

We evaluated the performance of the **Extra Tree**, **CAT Boost**& **XGBoost** algorithms, as we found out they are the best among those we studied.

a) XGBoost

It includes L1 and L2 regularization terms to prevent overfitting—a common issue in

regression tasks involving high-dimensional data like car specifications. XGBoost is a powerful ensemble learning algorithm that builds boosted decision trees to minimize prediction errors efficiently. It's ideal for used car price prediction due to its high accuracy, speed, and ability to handle complex, real-world datasets.

b) CAT Boost

It is a high-performance gradient boosting algorithm that handles categorical features natively without explicit encoding, making it ideal for used car price prediction where many features (like brand, model, fuel type) are categorical. It also prevents overfitting and performs well even with smaller datasets.

c) Extra Trees (Extremely Randomized Trees)

It is an ensemble learning method that builds multiple unpruned decision trees using random splits and averages their outputs to improve prediction accuracy. It is highly efficient for regression tasks like used car price prediction due to its speed and ability to reduce overfitting.

IV. Result & Discussion:

The **Extra Trees** (Extremely Randomized Trees) regression model proved to be very effective for predicting the target variable with an **accuracy of 89.02%.** Such accuracy indicates that the model also has a good generalization capability for unseen input. Likewise, the model produced an **R^2 score of 81.76%**, meaning that approximately **82% variance** in the dependent variable can be predicted from the independent features. **Mean Absolute Error** (MAE) at a value of **1.75** states that the average prediction errors are very low, while a **mean square error** at **26.54** indicates an average error deviation towards a moderate level, which also features a root **mean square error** of **5.15** for this model.

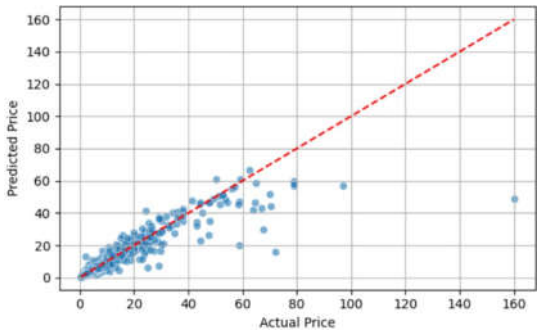

Figure 4.1 Actual vs Predicted prices using Extra Tree

The **confusion matrix of the Extra Trees** model reveals its finest performance in classification by identifying **570 Low** and **496 High** class instances accurately. It only **misclassifies 60 Low as High and 69 High as Low** instances, which translates to the model being very reliable. The darker diagonal blocks confirm the prediction rates that are ideal and accurate.

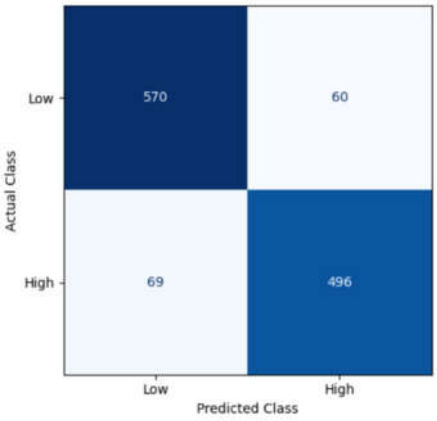

Figure 4.2 Confusion Matrix of Extra Tree

The regression model known as **XGBoost** (Extreme Gradient Boosting) came up with sound **predictions** at a value of **82.01%.** Though slightly inferior to the Extra Trees model, it does indicate a very good fit to the data. The model also boasts of an **R^2 score** of **78.41%,** meaning approximately **78%** of the **variation** in the target variable is explained by the features. It had a **Mean Absolute Error** (MAE) of **2.49**, which tended to be a bit higher on average than the Extra Trees. Likewise, the values of **Mean Squared Error** (MSE) and **Root Mean Squared Error** (RMSE) measured **31.43 and 5.60**, respectively, suggesting prediction errors at moderate levels.

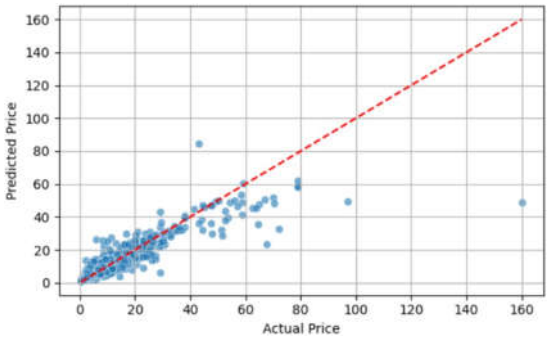

Figure 4.3 Actual vs Predicted prices using XGBoost

The **confusion matrix showing Extra Trees** classifier performance is very convincing with **570 Low and 496 High** instances correctly classified. There were only **60 Low misclassified** as **High and 69 High as Low**, demonstrating few errors. This is a validation of the high accuracy of the model and its ability to classify reliably.

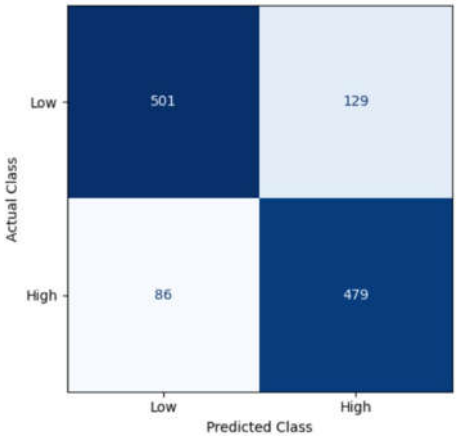

Figure 4. 4 Confusion Matrix of XGBoost

The **CAT Boost** regression model performed finely by yielding an **accuracy level** of **81.25%** which indicates a fairly strong fit on the data. The **R^2 score of 77.50%** indicates that the model could explain a good portion of the variance of the target variable. The model recorded an **MAE of 2.61**, **MSE of 32.75**, and **RMSE of 5.72,** showing it gave slightly more prediction errors as compared to Extra Trees and XGBoost, but still a reliable model, especially for handling categorical variables quite well.

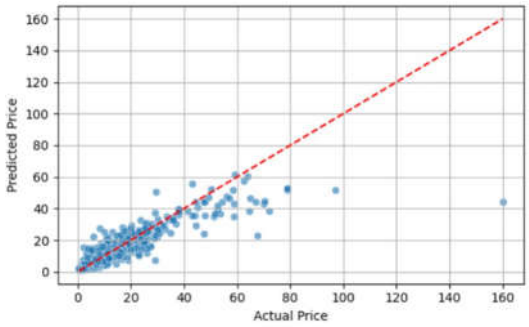

Figure 4.5 Actual vs Predicted prices using CAT Boost

The **CAT Boost model's confusion matrix** indicates that it was able to **correctly classify 482 instances** into the **Low class** and **489 into the High class.** However, **148 instances were misclassified as High instead of Low** while **76 were misclassified as Low instead of High**. The model seems to have performed poorly in a sense in predicting the Low predictions, but overall the performance of the model was balanced.

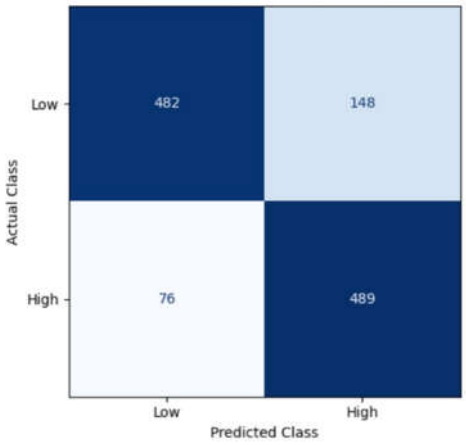

Figure 4.6 Confusion Matrix of CAT Boost

	Extratrees	XGBoost	CAT Boost
Accuracy	89.02%	82.01%	81.25%
R^2	81.76%	78.41%	77.50%
MAE	1.75	2.49	2.61
MSE	26.54	31.43	32.75
RMSE	5.15	5.60	5.72

V. CONCLUSION

This research work has presented a machine learning-based method for more accurate prediction of prices of used cars by analysing major features like brand, model, year of manufacture, mileage, fuel type, transmission type, and ownership history. Different supervised learning algorithms such as Linear Regression, Decision Tree, Random Forest, and Extra Trees Regressor have been used in this study and thus their performance has been compared in identifying the most effective for the regression task in question.

We maintained that after the application of essential preprocessing methods such as Data Cleaning, Feature Encoding, and Normalization, we then applied the training dataset upon model training. According to R^2 scores, Mean Squared Error, and visual comparisons of actual vs, predicted prices, attained accuracy and reliability models have been evaluated.

It demonstrated, however, that machine learning models could be effectively employed to augment a conventional price estimation in the used car market into a quantitative process, which would lead every agent in the transaction to make well-informed decisions without taking part in what he probably surmises or thinks and ultimately facilitate an appraisal, whether individual or collective, as well in housekeeping at the dealership level or online car resale platforms.

In future, incorporating real-time market trends, geographic data, and using deep-learning and joint-collaborative models-based techniques could enhance prediction accuracy. Adding further value, this model could also be extended into an easy-to-use web or mobile application.

V. FUTURE SCOPE

Machine learning has proven to be a great tool that has come so far in predicting used car prices with high reliability; however, these areas still open up to much more improvements and exploration. The first is to have a much more extensive and diversified data set in which the utility of the models might be increased as most of the current studies are based on limited or localized data. Using, for example, a network that has cross-country or aggregated industry level data, a more comprehensive market trend can be achieved.

Advanced feature engineering and automated feature-selection techniques may uncover hidden determinants, such as accident history, service records, and user ratings, that are significant constituents of pricing. Other factors in real time, such as market variations and seasonality, could also form the basis for dynamic pricing models.

Ensemble methods and advanced deep learning architectures, especially those that involve the combination of convolutional and recurrent networks, will be able to capture even the most complex nonlinear forms and temporal trends that could lead to more accurate predictions. Moreover, XAI techniques can address the tension created by the black box of some of the models to allow for credibility and trust without restricting insight on critical predictive features.

Further promising avenues of exploration may include flexible online machine learning-based pricing systems that provide car dealerships, reseller platforms, and individual consumers with real-time valuation applications. Add social community features and make it mobile-friendly for user engagement.

In summary, traction can be gained in used car price prediction, through enriched data sets, advanced modeling techniques, explainability and innovations directed toward deployment.

References

[1] M. G. S. P. Pattabiraman Venkatasubbu, “Used Cars Price Prediction using Supervised Learning Techniques,”

International Journal of Computer Applications, vol. 9, no. 1S3, p. 9, 2019.

[2] Vanpariya and Harikrushna, “Using Different Machine Learning Techniques for Predicting the Price of Used Cars,” *International Journal for Scientific Research & Development (IJSRD),,* vol. 06, no. 10,2018, p. 5, 2018.

[3] C. Jin, “Price Prediction of Used Cars Using Machine Learning,” in *IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, Chongqing, 2021.

[4] A. R, “CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES,” *International Research Journal of Modernization in Engineering Technology and Science,* vol. 04, no. 2february2022, p. 6, 2022.

[5] R. R. Patil, “Used Car Price Prediction Using ML,” Changu Kana Thakur Arts, Commerce and Science College, Panvel, Panvel, 2024.

[6] A. S. Pillai, “A Deep Learning Approach for Used Car Price Prediction,” *Journal of Science & Technology By The Science Brigade (Publishing) Group,* vol. 3, no. 3, p. 22, 2022.

[7] A. Nikhade and R. Borde, “Car Price Prediction using Machine Learning,” *International Advanced Research Journal in Science, Engineering and Technology,* vol. 9, no. 4, April 2022, p. 6, 2022.

[8] E. Pandit, H. Parekh , P. Pashte and Aaka, “Prediction of Used Car Prices using Machine Learning Techniques,” *International Research Journal of Engineering and Technology,* vol. 09, no. 12, p. 6, 2022.

[9] S. Tamrakar, V. Dutta, A. Singh, K. Sahu and A. Dewangan, “Car Price Prediction Using Data Science and Machine Learning,” *International Research Journal of Modernization in Engineering, Technology and Science, Volume:06/Issue:05/May-2024,* vol. 6, no. 5, p. 22, 2024.

[10] AlShared and Abdulla, “"Used Cars Price Prediction and Valuation using Data Mining Techniques",” "Department of Graduate Programs & Research, Dubai, 2021.

[11] H. Ahaggach, L. Abrouk, S. Foufou and E. Lebon, “"Predicting Car Sale Time with Data Analytics and,” Lifecycle Management. PLM in Transition Times: The Place of Humans and Transformative Technologies, France, 2023.

[12] A. Chandak, P. Ganorkar, S. Sharma, A. Bagmar and S. Tiwari, “Car Price Prediction Using Machine Learning,” *International Journal of Computer Sciences and Engineering,* vol. 7, no. 5, p. 8, 2019.

[13] K.Samruddhi and D. R. Kumar, “Used Car Price Prediction using K-Nearest Neighbour Based Model,” *International Journal of Innovative Research in Applied Sciences and Engineering,* vol. 4, no. 2, p. 4, 2020.

[14] L. Bukvić, J. P. Škrinjar, T. Fratrović and B. Abramović, “Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning,” *Sustainability,* p. 17, 19 December 2022.

[15] M. Hankar, M. Birjali and A. Beni-Hssane, “Used Car Price Prediction using Machine Learning,” in *11th International Symposium on Signal, Image, Video and Communications (ISIVC)*, El Jadida, 2022.

[16] L. P. Mudarakola, D. S. Prakash, K. L. N. Shashidhar and D. Yaswanth, “Car Price

Prediction Using Machine Learning," *International Journal for Research in Applied Science & Engineering Technology (IJRASET),* vol. 12, no. V May 2024, p. 9, 2024.

[17] N. Pal, P. Arora, D. Sundararaman, P. Kohli and S. S. Palakurthy, "How much is my car worth? A methodology for predicting used cars prices using Random Forest," in *Future of Information and Communications Conference (FICC) 2018*, India, USA, 2017.

[18] N. Shanti, A. Assi, H. Shakhshir and A. Salman, "Machine Learning-Powered Mobile App for Predicting Used Car Prices," in *ACM Symposium on Document Engineering*, Limerick, 2021.

[19] S. Muti and K. Yıldız, "Using Linear Regression for Used Car Price Prediction," *nternational Journal of Computational and Experimental Science and Engineering (IJCESEN),* vol. 9, no. 2023, p. 6, 2023.

[20] V. K. H, S. V, A. V and S. Srinivas, "Comparative Analysis of Machine Learning Algorithms for Used Car Price Prediction," *International Journal of Current Science Research and Review,* vol. 07, no. 09 September 2024, p. 9, 2024.