IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141036

CarPrice Prediction Using Machine Learning

Nikhil Dnyaneshwar Bagul[1], Kaustubh Bhave[2], Manoj V. Nikum*[3]

Student Of MCA, SJRIT Dondaicha, KBC NMU Jalgaon, Maharashtra, India[1]

Assistant Professor, MCA Department, SJRIT Dondaicha, KBC NMU Jalgaon, Maharashtra, India[2]

Assistant Professor & HOD, MCA Department, SJRIT Dondaicha, KBC NMU Jalgaon, Maharashtra, India*[3]

Abstract: Correct car price prediction is significant in automotive manufacturing as it offers important benefits to the producers, dealers, customer by supporting honest pricing, inventory management, and informed management. This research develops a machine learning system that predicts the prices of cars based on an alteration of attributes, like name, price, model, year, km driven, and type of fuel. The work used a large dataset; the dataset was filtered so that missing values, data inconsistencies, and outliers in the data were reduced. like Linear Regression, Decision Trees, and Random Forests are working to make predictive models. The presentation of these models is evaluated using metrics like R-squared R^2 for accuracy and reliability. The outputs display the ability of Machine Learning techniques to deliver more accuracy in car price predictions,s howing their practical stability in the automotive domain. By dealing with issues like data quality, feature selection, and model interpretability, this study offers a solid basis for developing similar predictive systems. Besides, the study suggests that improvement, including incorporating real-time market data, can be considered to increase the accuracy of the prediction. This study has made it clear that Machine Learning plays a very complex role in changing the pricing strategies and supporting the stakeholders in driving an active automotive market.

Keywords: Car price prediction, machine learning, Automotive Manufacturing, predictive modeling, Linear Regression, data preprocessing, feature selection, R-squared, pricing strategies, integration, model Real-time Market Data, inventory Management

I. INTRODUCTION

Automotive production has been knowledgeable remarkable growth in the past few years, sustained by increasing demand for both new and used vehicles. It produced a highly powerful market, where predicting the cost of a car is a significant task for many stakeholders in this market, including the producers, dealers, and customers. Correct cost estimation allows consumers and suppliers to make informed decisions, promotes transparency in transactions, and helps businesses in strategic -making.choice However, the decision regarding the cost of a car involves a comprehensive analysis of various factors, such as the name of the vehicle, type of fuel, model, year, km, and prevailing market trends. Machine Learning has emerged as a creative technology to solve complex tasks of prediction, such as car price estimation. The machine learning models are superior when it comes to handling high volume data and identifying patterns which traditional methods often tend to avoid. Using historical data, ML can give right and efficient price predictions. It is this knowledge that aids buyers and sellers in setting a fair price and can also help businesses by bringing better inventory management, high customer approval, and other better market strategies.

This paper focuses on the development of a machine learning-based system that predicts car prices. The authors use a dataset of many vehicle attributes and preprocess data to remove missing values, outliers, and unpredictability. Machine learning algorithms such as Decision Trees, Random Forests, and Linear Regression make it possible for the system to predict models. It analyses various models for accuracy, reliability, and accuracy in regard to metrics such as R-squared and Mean Absolute Error. The system will be flexible and scalable as well, making it able to be used by people in different markets and places.

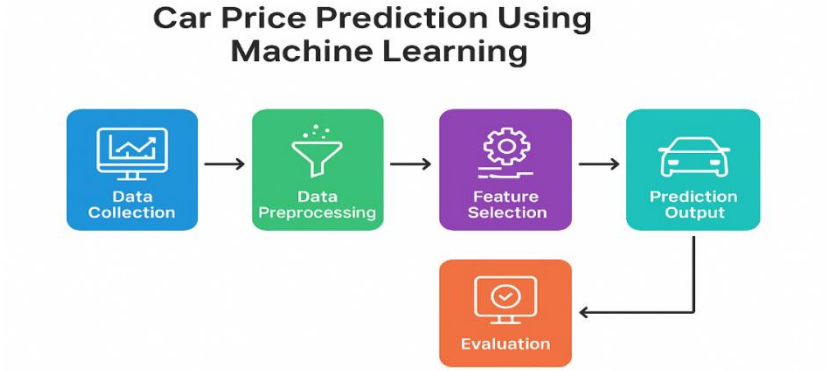
This paper also looks into future improvements to include ensemble learning techniques with real-time market data in the combination and applying the explainable AI frameworks on the system to make the built structure more transparent and trustworthy. It showcases how Machine Learning can change pricing strategies and assist stakeholders in driving innovation in the fast-changing world of the automotive market. Machine Learning has huge scope not just to evolve existing processes, but new scopes for developing an industry for growth and efficiencies of a really competitive industry.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141036



II. LITERATURE SURVEY

In[2]This study is designed to predict used car prices thr ough machine learning.the author has performed data preprocessing and tested several regression algorithms. the decision tree algorithm showed the best performance obtaining the R2 score of 0.95, lowest RMSE, and MSE which indicated high accuracy with minimum errors in prediction.

In[6] the study focuses on using data-driven methods of machine learning to help predict automobile prices, helping both buyers and sellers to make well-informed decisions.By evaluating factors like market trends and brand reputation, ML models offer accurate, flexible pricing, minimizing the risks of overpricing or underpricing. While they increase transparency and trust, model accuracy depends on data quality and external factors like economic changes.

In[3] the study focuses on Predicting car prices is challenging because of the multiple factors influencing them. This study highlights data preprocessing and employs an ensemble method that includes multiple algorithms, resulting in a remarkable accuracy of 92.38%. while the method needs greater computational resources, it delivers exceptional predictive accuracy for automotive market analysis.

In[1] the study focuses on using AI techniques and the Kaggle dataset and applying extra tree regressor and random forest algorithms to predict price of a used car. This test dataset,produced by picking values at random from the original dataset, is used to assess the model's predictions. After thorough analysis, it is concluded that both Extra Trees Regressor and Random Forest are impressive decisions regarding regression tasks,delivering extremely correct predictions despite dataset size.

In[5] The study develops a machine learning model to predict used car prices, with a focus on data preprocessing.The model is implemented as a web application,helping buyers and sellers make well informed decisions and highlighting the capabilities of machine learning in the automotive industry.

In[7] the study focuses on emphasizes the significance of accurate predictions in the increasing used car market, analyzing machine learning models for predicting used car prices.It provides a comparative analysis of algorithms from eight research papers, describing their strengths and weaknesses.

In[4] the study focuses on supervised machine learning techniques for used car price prediction with exacting validation. the linear regression developed as the most correct model after the preprocessing and feature extraction. Particularly utilizing key attributes like price and model. This research paper highlights optimizing the pricing using data driven methods,minimizing errors,and assisting consumers to well informed choices.

In[3] the study focuses on As the prices of new cars continue to increase, the demand for used car sales, especially at the Taluka level, is growing. An automobile price prediction system is necessary to estimate the price of used cars depending on several attributes to address this. Using insights from survey papers and the linear regression algorithm, this model can provide accurate price estimates, with the potential to make a userfriendly UI application.

III. RESEARCH METHODOLOGY

The methodology part describes how the car price prediction model was developed.It explains the steps collected to arrange the data, such as cleaning it and selecting the Significant features. It additionally discussions regarding reason

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141036

the linear regression model was selected and In the way that it was trained and tested.At last, it describes how the model's execution was Calculated to view how Properly it predicts car prices.The goal is to show the process evidently so others can understand or repeat it.

Selecting the right Technique for car price prediction is key to accepting correct results. unique methods task better with different types of data. A good method gives more secure price estimates, which helps car customers and dealers. It must fit the data and be easy to know. selecting the incorrect method can guide to poor predictions. It's as well important to regard how much time and computer power the method needs, especially for large-scale use. The right selection confirms the model works properly and is easy to apply.

I Data Collection:
The dataset originates out of the Faster website and include knowledge about used items. It includes information like the product name, company, year, kilometers driven, fuel type (e.g., petrol, diesel), and price.The dataset has 892 records over 6 different product types.

II Data Preprocessing:
Data preprocessing is an important phase to ready your data for machine learning. It supports confirm the data equals in the correct layout for the model to task successfully. Here are the main tasks involved:

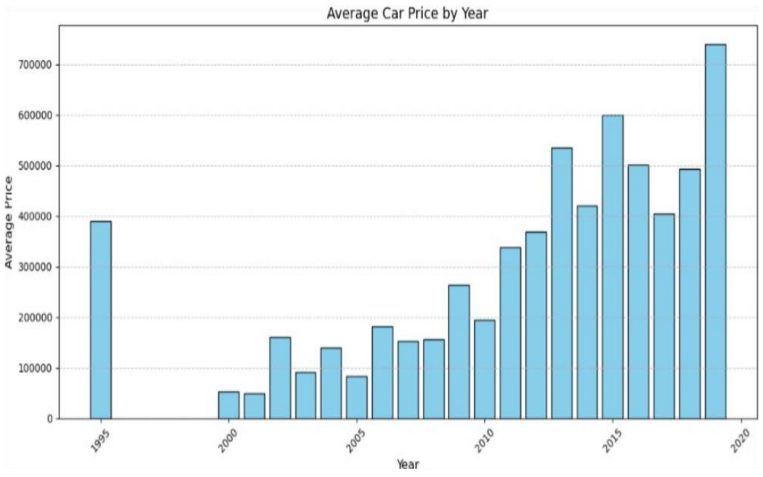
A. Converting Object to Integer:
Some columns might have data in text format (like categories). These requirement become converted into numbers. For categories, you can utilize methods like one-hot encoding (creating new columns for per category) or label encoding (giving each category a unique number). For organized categories (like low, medium, high), you can assign numbers located on their order.

B. Removing Unwanted Data:
Recognize and delete columns that don't support by predictions, like IDs or irrelevant information. If there are lost values (null spaces), you can any overflow them in with a value (like the average or median) or delete the rows/columns with lost data.

C. Handling Outliers:
Outliers are values that are truly unique from the repose of the data. You can detect them applying methods like Z-scores or box plots. Once recognized, outliers can be deleted, changed (like using a log scale), or restricted to cut their impact on the model.

D. Converting Data Types:
Confirm the numbers are in the accurate format. Like, if a number is displayed as a decimal (float) still must be an integer, update it. This confirms the data is suitable with the machine learning model you're making use of. In brief, data preprocessing clears and changes data to create it prepared for machine learning.

III Data Analysis:
Average Car Prices Over the Years



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

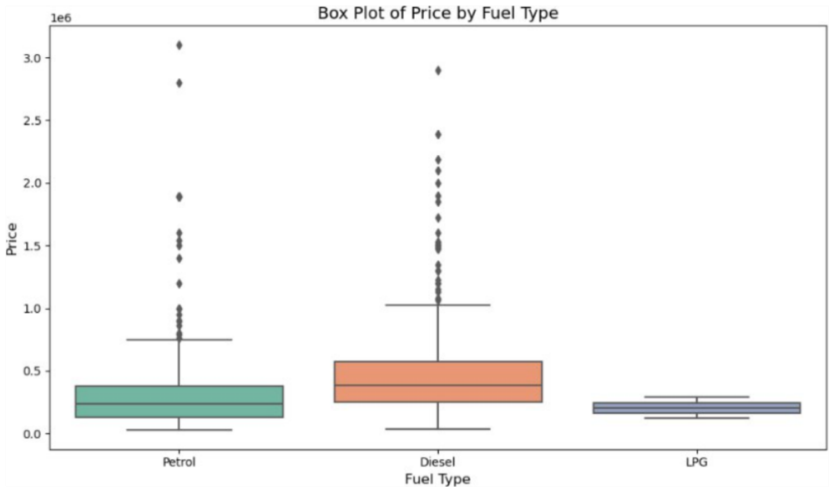
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141036

The bar chart titled Average Car Price by Year shows the detailed trend of changes in car prices over 25 years from 1995 to 2020. The x-axis is the year and the y-axis represents the average car price market preference. It clearly shows a peak in 1995, with an average price that was quite high compared to the subsequent years, which might indicate an anomaly or a factor specific to the data. Average prices from 2000 onward show a steep trend of increase, except after 2010 they show a more pronounced graph. This demonstrates that over the years, the price rise for cars has been consecutive and the highest average was attained in 2020. However, the graph shows factors in the economy and wider markets that have affected these prices, such as the increase in inflation, newer automobile technology, and varied consumer demand. The crucial input variable for the car price prediction model is the year of production, as is evident by visualizing the trend because it directly goes with the rising price trends observed year after year.

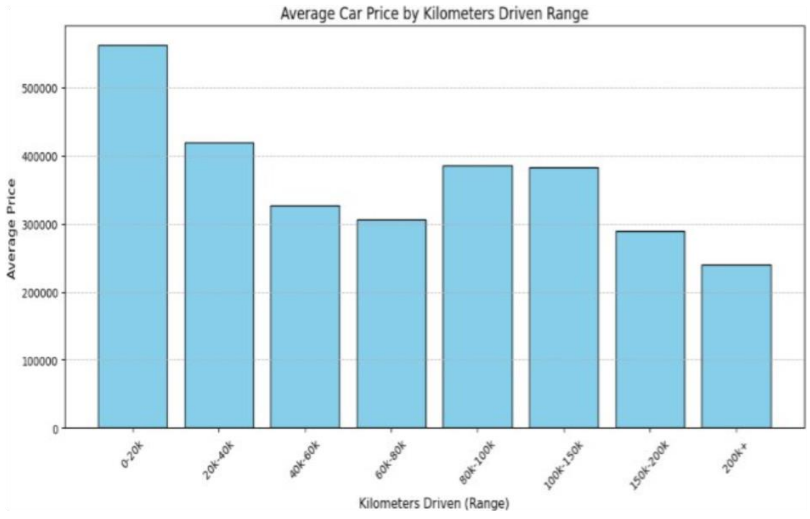
Box Plot of Price by Fuel Type



The box plot "Box Plot of Price by Fuel Type" contains the comparative price analysis of cars, taken in three types: Petrol, Diesel, and LPG. The horizontal scale or x-axis represents the types, while the vertical or y-axis represents the price for cars. Each box plot represents the distribution of car prices for one type of fuel, showing, among other things, the median in the form of a line inside the box, interquartile range as the height of the box, and data spread through whiskers and outliers.

The plot shows that Diesel cars tend to be priced higher than Petrol and LPG cars, with the median and interquartile range being higher for Diesel cars. Petrol cars have a slightly wider price range with greater variability, having a number of outliers that depict cars that are priced much higher than the range. LPG cars have the lowest price range with minimal variability and few outliers.

Average Car Price by Kilometers Driven Range:



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025

DOI: 10.17148/IJARCCE.2025.141036

The bar chart titled "Average Car Price by Kilometers Driven Range" shows how the range of kilometers that a car has been driven relates to its average price. The xaxis groups the cars into mileage ranges (0–20k, 20k– 40k, etc.) and the y-axis represents the average price in each range.

From the chart, it is obvious that there is an inverse relationship between the kilometers driven and the price of the car. Cars falling into the 0–20k range have an average price that is more considerable, which shows higher valuation in the market for low-mileage vehicles. With increasing mileage counts, the average price continues dropping and is more of the devaluation of a vehicle's price with higher usages. It levels off somewhat as the middle mileage ranges fall at around 80k–150k before it falls and takes up the lowest average value at 200k+.

V Feature Engineering:

Feature engineering is about creating updated features or modifying present ones to build a model developed at awareness patterns. Suppose, Alternatively of applying only the manufacturing year of a car, you can calculate its "age" by removing the manufacturing year from the present year. This "age" feature can provide the model additional meaningful information, supporting it create more effective predictions on the car's price.

VI Model Selection:

1.Linear Regression:

This model looks at the connection between car properties (like make, model, year, mileage) and the price. It develops a easy straight-line equation to predict the price. It's simple to get, and the model will display
how every feature (like age or mileage) impacts the price
.

2.Random Forest:

This model builds multiple decision trees, where each tree tries to predict the price applying separate features. It then joins the predictions of the trees to get a last result. Random Forest can manage complicated connection in the data and is improved at predicting prices when models are not easy.

3.R2 Score:

The R2 score informs us how well the model discribes the differences in car prices. An R2 score of 1 method the model perfectly predicts prices. A larger score is improved, expressing that the model does a excellent job at explaining the price changes.

IV. RESULTS AND DISCUSSION

Performance evaluation was done using the R^2 score, which depicts how much variation in the dependent variable could be explained by independent variables. The results of all the models are compiled below in the table:

Model	R2 Score
Linear Regression	0.89
Random Forest	0.84
Hybrid model	0.86

The Linear Regression model had the highest R^2 score at 0.89. This shows that it explains 89% of the variability in automobile prices based on the input features. Its high performance shows that the relationship between the features and the target variable is largely linear, which makes this model a good choice for this dataset.
It scores an R^2 of about 0.84, hence not as great as in the case of linear regression. Though typically robust, Random Forest also does best with highly complex relationships or noisy data types, neither of which are seen much in these datasets: one lacking extreme nonlinearity or noisy patterns in which Random Forest usually excels.
A Hybrid model could average predictions from both Linear Regression and Random Forest the highest R^2 score at 0.86 balancing the bias of linear models with the flexibility of Random Forest.

V. CONCLUSION AND FUTURE SCOPE

The study shows that the Linear Regression model outperforms the Random Forest model in predicting automobile prices for the given dataset. The Linear Regression model achieved the highest R^2 score of 0.89, indicating that it can explain 89% of the variability in car prices based on the input features.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 10, October 2025
DOI: 10.17148/IJARCCE.2025.141036

This suggests that the dataset primarily exhibits a linear relationship between the features and the target variable, making Linear Regression the most suitable model for this scenario.

The Random Forest model, with an R^2 of 0.84, also performed reasonably well but did not surpass Linear Regression, likely because the dataset lacks strong nonlinearity or noise where Random Forest typically excels.
Interestingly, the Hybrid model, which combines predictions from both Linear Regression and Random Forest, achieved an R^2 score of 0.86. This demonstrates that integrating linear and ensemble learning approaches can yield a balanced performance, leveraging the strengths of both models.

For future work, more advanced ensemble techniques such as Gradient Boosting or XGBoost can be explored to further enhance prediction accuracy. Additionally, expanding the dataset with more diverse and nonlinear features may help improve model generalization and reveal deeper insights into automobile price determinants.

REFERENCES

[1]. R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts," *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 404–411, 2004.
[2]. H. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, 1958.
[3]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[4]. A. See, P. J. Liu, and C. D. Manning, "Get to the Point: Summarization with Pointer-Generator Networks,"*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1073–1083, 2017.
[5]. A. Vaswani et al., "Attention is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.
[6]. C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research (JMLR)*, vol. 21, no. 140, pp. 1–67, 2020.
[7]. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.
[8]. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
[9]. Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 8, pp. 328–341, 2020.
[10]. T. Li, Q. Li, and H. Lin, "Multimodal Document Summarization with Visual and Textual Attention,"*Proceedings of the 2022 Conference on Computational Linguistics*, pp. 201–212, 2022.
[11]. A. Wierman, Z. Liu, I. Liu, and H. Mohsenian-Rad, "Opportunities and Challenges for Data Center Demand Response," *Proceedings of the International Green Computing Conference*, vol. 7, pp. 1–10, 2014.
[12]. N. Hogade, S. Pasricha, and H. J. Siegel, "Energy and Network Aware Workload Management for Geographically Distributed Data Centers," *IEEE Transactions on Sustainable Computing*, vol. 7, no. 2, pp. 400–413, 2021.
[13]. D. G. Feitelson, D. Tsafrir, and D. Krakov, "Experience with Using the Parallel Workloads Archive," *Journal of Parallel and Distributed Computing*, vol. 74, no. 3, pp. 2967–2982, 2014.
[14]. B. McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data,"*Artificial Intelligence and Statistics (AISTATS)*, Proc. PMLR, vol. 10, pp. 1273–1282, 2017.
[15]. F. Van den Abeele, J. Hoebeke, G. Ketema, I. Moerman, and P. Demeester, "Sensor Function Virtualization to Support Distributed Intelligence in the Internet of Things," *Wireless Personal Communications*, vol. 81, no. 4, pp. 14– 18, 2015.

