IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

# Car Sales Prediction Using Machine Learning

**Ketan Kumar*1, Mayank Kumar*2, Prof. Sunil Kumar Chowdhary*3,**
*1,2,3 School of Computing Science & Engg. Galgotias University Greater Noida, Uttar Pradesh-203201

***Abstract***— ***Sales prediction is the current trend in which all the businesses thrive, and it also aids the concern in determining the future goals for it and its plan and procedure to achieve it. Sales of cars do not contain any independent variable since various factors such as horsepower, model, width, fuel type, height, price, city-mileage, highway-mileage and manufacturer are the various features that influence the sales. The primary problem I want to solve with my project, "Car Sales Prediction," is the challenge of accurately forecasting car sales in an ever-changing market. Manufacturers, dealerships, and investors often struggle with predicting sales due to that influence the market. By addressing this issue, the project aims to provide a more accurate model for predicting car sales, helping stakeholders make better-informed decisions and develop effective marketing strategies. In the existing body of work on car sales prediction, several gaps remain that my project aims to address. traditional prediction models. My project aims to fill these gaps by developing a more prediction model that real-time data and advanced machine learning algorithms. The results of the "Car Sales Prediction" study are Cost Reduction, Customer Satisfaction, Innovation and Improvement. I want to address the problem of car sales prediction for several reasons. Technological Advancement, Consumer Benefits, Sustainability, Personal and Growth.***

## 1. Introduction

Cars play a crucial role in our daily lives and are a big part of the global economy. Buying and selling cars is a common activity, and knowing the right price is very important for everyone involved. Predicting car prices accurately helps buyers get a fair deal and allows sellers to set competitive prices. This is especially important now because more people are using online platforms to buy and sell cars.

This paper focuses on using machine learning to predict car prices. By analyzing information like the car's brand, year of manufacture, mileage, engine size, and condition, our system can suggest a reliable price for the car. This is useful for car dealers, individual buyers, and sellers who want to make informed decisions.

The system also includes a web interface that is easy to use. Users can enter details about a car, such as its age, mileage, and fuel type, to get a price prediction. The web interface also allows users to download a report with all the details. By combining advanced technology with a simple design, this project makes price prediction accessible to everyone.

Overall, this paper helps to solve the problem of price uncertainty in the car market. It uses data and technology to provide accurate and quick results, making it a helpful tool for anyone involved in buying or selling cars.

## 2. Literature Survey

Kuiper, S. (2008) proposed a multivariate regression model for classifying and predicting numeric values, showcasing its application in forecasting the price of 2005 General Motors (GM) vehicles. The study highlighted that predicting car prices does not necessitate specialized knowledge; publicly available data suffices. Kuiper demonstrated variable selection techniques to identify the most relevant variables for inclusion in the model, enabling more precise predictions.

Pal, N et al. (2019) introduced a methodology utilizing Random Forest for predicting used car prices. Their research, conducted with a Kaggle dataset, achieved an accuracy of 83.62% for test data and 95% for training data. Key features like price, kilometers driven, brand, and vehicle type were identified as significant predictors after removing outliers and irrelevant features. The study emphasized that Random Forest's sophistication ensured superior accuracy compared to earlier models.

Praful Rane, Deep Pandya, and Dhawal Kotak employed regression algorithms such as Lasso, Linear, and Ridge Regression to forecast the exact cost of automobiles, instead of categorizing them into price ranges. This approach facilitated precise cost predictions. Additionally, the authors developed a user-friendly interface allowing users to input details and receive accurate car price predictions.

Laveena D'Costa et al. applied machine learning algorithms to estimate the true value of cars when sold to dealers. Their methodology employed a multiple linear regression model, dividing the dataset into training and testing subsets. The study underlined the significance of predicting used car prices, especially for vehicles not sourced directly from manufacturers.

In summary, the literature on car sales prediction demonstrates an evolution from traditional statistical approaches to advanced machine learning methodologies. Traditional models provide a foundational understanding, while machine learning enhances predictive accuracy through feature engineering and real-time data integration. Future studies could focus on leveraging real-time data and exploring additional features to refine predictive models further.

## 3. Methodology

The methodology for developing the car sales prediction model consists of the following steps:

### 3.1. Dataset Overview

The Cardetails.csv dataset includes various attributes necessary for car price prediction. These attributes are:
Name: The brand and model of the car.
Year: Manufacturing year of the car.
Kilometers Driven: Distance the car has been driven.
Seller Type: Indicates whether the seller is an individual or a dealer.
Transmission Type: Type of transmission (Manual or Automatic).
Owner: Number of previous owners.
Mileage: Fuel efficiency in kmpl.
Engine: Engine capacity in CC.
Max Power: Maximum power output in bhp.
Seats: Number of seats in the car.

### 3.2. Data Preprocessing

To ensure accurate predictions, the data was cleaned and transformed as follows:
Handling Missing Values: Imputed missing data to ensure completeness.
Feature Engineering: Extracted brand names from the Name column.
Categorical Encoding: Converted categorical variables into numeric formats using label encoding.
Scaling: Standardized numerical features like mileage, engine, and max power for uniformity.

### 3.3. Model Training

Algorithm Selection: The Random Forest Regressor was chosen for its robustness and ability to handle complex interactions between features.

Training Process: The dataset was split into training and testing sets (80:20 ratio) to evaluate model performance. The model was trained using the training set and validated on the test set.

Hyperparameter Tuning: Parameters such as the number of trees and maximum depth were optimized to enhance model accuracy.

### 3.4. Prediction and Adjustment

The model predicts the car price based on input features.
Adjustments are made to the predicted price based on additional factors like car condition and color:
Condition Multiplier: Adjusts the price for conditions such as Excellent, Good, Average, and Poor.
Color Adjustment: Adds a premium for popular colors like White, Black, and Red.

### 3.5. User Interface Development

Using Streamlit, a web application was designed to:
Allow users to upload car images for visual representation.
Input car features via dropdowns, sliders, and text fields.

Display predicted prices dynamically.
Generate a PDF report summarizing the prediction.

### 3.6. Evaluation Metrics

The model's performance was evaluated using:
R-Squared Value: Measures the goodness of fit.
Mean Absolute Error (MAE): Indicates average prediction error.
Cross-Validation: Ensures consistent performance across different data splits

## 4. PROPOSED DESIGN

This project proposes a system that integrates a machine learning model with an interactive web interface. The design focuses on usability and accuracy, ensuring that users can easily access and understand predictions.

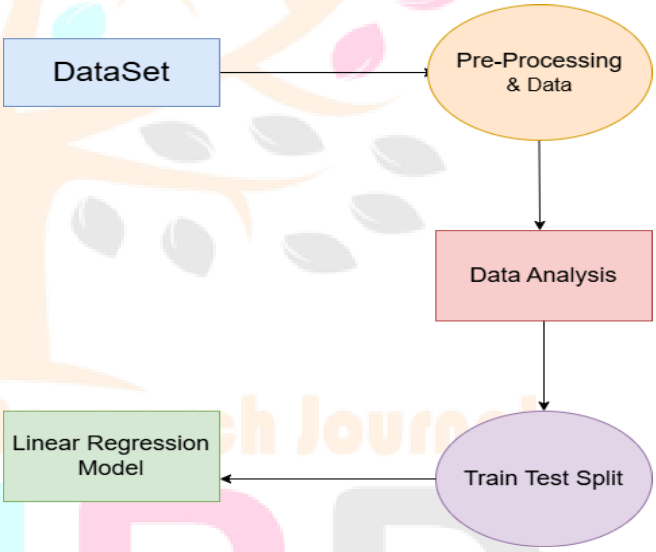
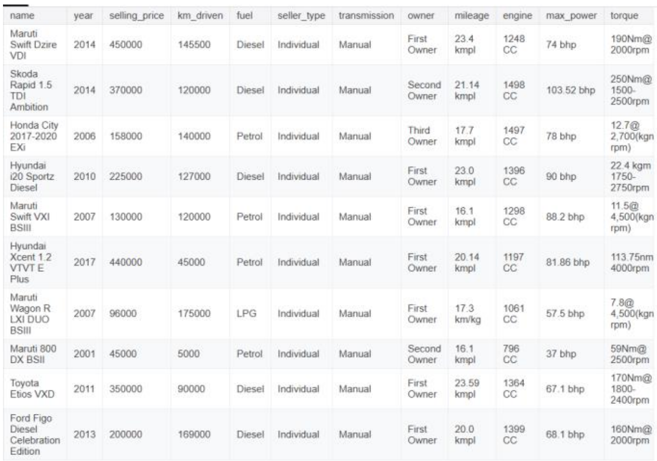


Figure 4.1. Linear Regression Model

(A). Data Collection and Preprocessing: The system uses a dataset containing various car attributes like brand, year, mileage, fuel type, and condition. Preprocessing steps include handling missing values, normalizing numerical features, and encoding categorical data.



| name | year | selling_price | km_driven | fuel | seller_type | transmission | owner | mileage | engine | max_power | torque |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Maruti Swift Dzire VDI | 2014 | 450000 | 145500 | Diesel | Individual | Manual | First Owner | 23.4 kmpl | 1248 CC | 74 bhp | 190Nm@ 2000rpm |
| Skoda Rapid 1.5 TDI Ambition | 2014 | 370000 | 120000 | Diesel | Individual | Manual | Second Owner | 21.14 kmpl | 1498 CC | 103.52 bhp | 250Nm@ 1500-2500rpm |
| Honda City 2017-2020 EXi | 2006 | 158000 | 140000 | Petrol | Individual | Manual | Third Owner | 17.7 kmpl | 1497 CC | 78 bhp | 12.7@ 2,700(kgm rpm) |
| Hyundai i20 Sportz Diesel | 2010 | 225000 | 127000 | Diesel | Individual | Manual | First Owner | 23.0 kmpl | 1396 CC | 90 bhp | 22.4 kgm 1750-2750rpm |
| Maruti Swift VXI BSIII | 2007 | 130000 | 120000 | Petrol | Individual | Manual | First Owner | 16.1 kmpl | 1298 CC | 88.2 bhp | 11.5@ 4,500(kgm rpm) |
| Hyundai Xcent 1.2 VTVT E Plus | 2017 | 440000 | 45000 | Petrol | Individual | Manual | First Owner | 20.14 kmpl | 1197 CC | 81.86 bhp | 113.75nm 4000rpm |
| Maruti Wagon R LXI DUO BSIII | 2007 | 96000 | 175000 | LPG | Individual | Manual | First Owner | 17.3 km/kg | 1061 CC | 57.5 bhp | 7.8@ 4,500(kgm rpm) |
| Maruti 800 DX BSII | 2001 | 45000 | 5000 | Petrol | Individual | Manual | Second Owner | 16.1 kmpl | 796 CC | 37 bhp | 59Nm@ 2500rpm |
| Toyota Etios VXD | 2011 | 350000 | 90000 | Diesel | Individual | Manual | First Owner | 23.59 kmpl | 1364 CC | 67.1 bhp | 170Nm@ 1800-2400rpm |
| Ford Figo Diesel Celebration Edition | 2013 | 200000 | 169000 | Diesel | Individual | Manual | First Owner | 20.0 kmpl | 1399 CC | 68.1 bhp | 160Nm@ 2000rpm |

Figure 4.2. Dataset

(B). Feature Engineering: Features like mileage, engine size, and age are crucial predictors. Additional dynamic features such as market-adjusted prices based on condition and color are included to enhance accuracy.

(C). Model Selection: A Linear Regression model is chosen for its ability to handle both categorical and numerical data. It is trained on the preprocessed dataset to predict car prices.

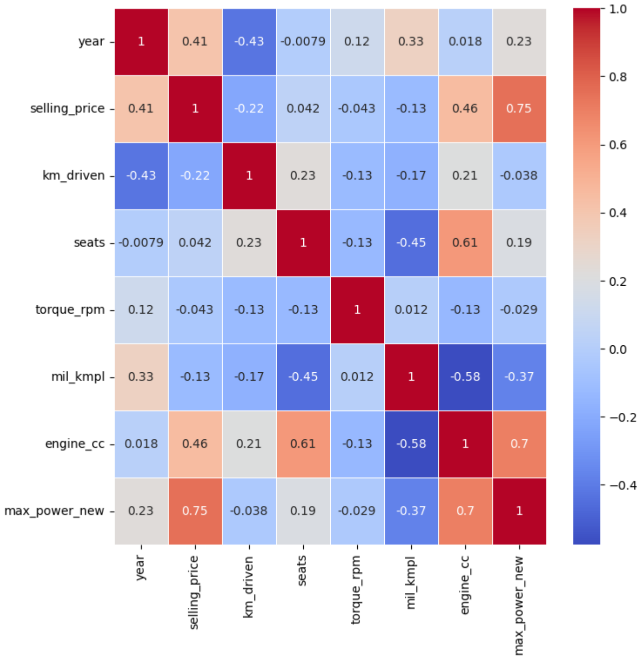


Figure 4.3. Data Visualization With HeatMap

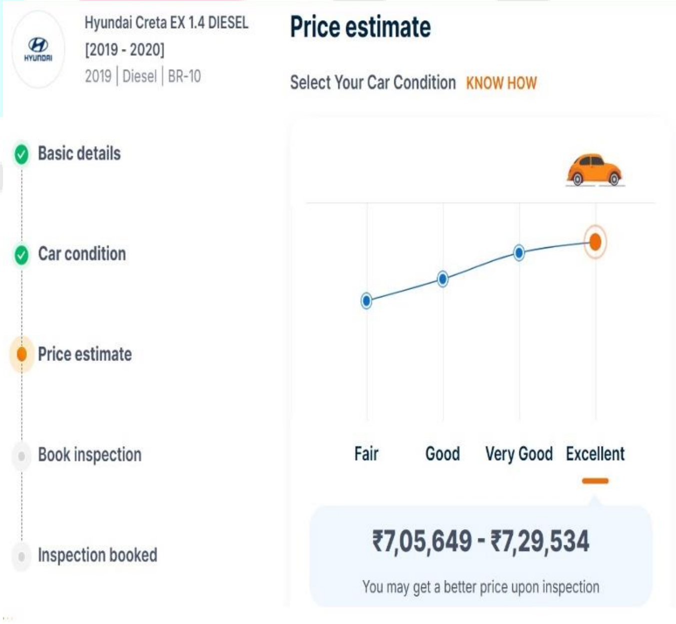
(D). Web Interface: The web interface is an indispensably portion of the car deals forecast framework, outlined utilizing Streamlit to guarantee user-friendly interaction. This interface serves as the essential point of interaction for clients, giving an instinctive stage to input car-related points of interest through direct shapes. Clients can indicate properties such as the car brand, year of fabricate, kilometers driven, fuel sort, and more. Upon accommodation, the framework forms the input and gives moment cost forecasts, dispensing with any complexity for the client. To upgrade the client involvement, the interface incorporates a highlight to create a nitty gritty PDF report. This report typifies the expectation and other important information, making it helpful for clients to spare or share the results.

(E). Dynamic Alterations: One of the standout highlights of the framework is its capacity to powerfully alter expectations based on real-world components. The framework considers user-provided conditions, such as the car's color and in general condition, to refine the anticipated cost. These alterations are grounded in showcase patterns and mimic how such variables might impact a vehicle's esteem in real-life scenarios. For occurrence, prevalent colors or cars in fabulous condition may get a higher anticipated cost, whereas less alluring properties may lead to a decrease. This highlight guarantees that the forecasts are not as it were data-driven but too commonsense and significant to showcase realities.

(F). Backend Advances: The backend of the framework is fueled by a strong combination of Python libraries. Scikit-learn is utilized for building and preparing the machine learning show, guaranteeing exact and proficient forecasts. Pandas and NumPy play a vital part in information dealing with and preprocessing, empowering consistent control and examination of the dataset. Moreover, ReportLab is utilized for creating proficient PDF reports, giving clients with a cleaned archive summarizing the expectation and related points of interest. This comprehensive utilize of backend advances guarantees that the framework is both effective and solid, able of dealing with complex information and conveying high-quality comes about.

## 5. Result and Discussion

Comparative Analysis of Price Prediction Models for a Used Vehicle. This study presents a comparative analysis between two car price prediction models: the commercial Car24 model and a custom-built predictive model. Both models were applied to the same vehicle dataset, which describes a Hyundai Creta EX 1.4 Diesel (2019) with manual transmission, first ownership, and an excellent condition rating. Key specifications such as mileage (20 kmpl), engine capacity (1376 CC), and maximum power output (83 bhp) were consistent across both evaluations.



Hyundai Creta EX 1.4 DIESEL [2019 - 2020]
2019 | Diesel | BR-10

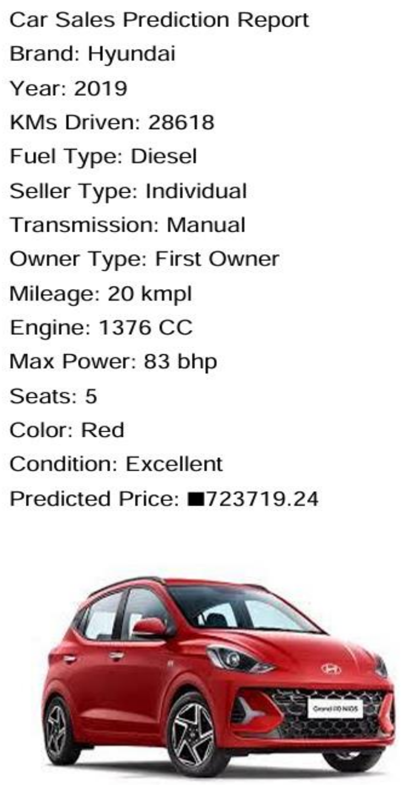
Price estimate

Select Your Car Condition KNOW HOW

₹7,05,649 - ₹7,29,534

You may get a better price upon inspection

Figure 5.1. car24 Prediction



Car Sales Prediction Report

Brand: Hyundai

Year: 2019

KMs Driven: 28618

Fuel Type: Diesel

Seller Type: Individual

Transmission: Manual

Owner Type: First Owner

Mileage: 20 kmpl

Engine: 1376 CC

Max Power: 83 bhp

Seats: 5

Color: Red

Condition: Excellent

Predicted Price: ■723719.24

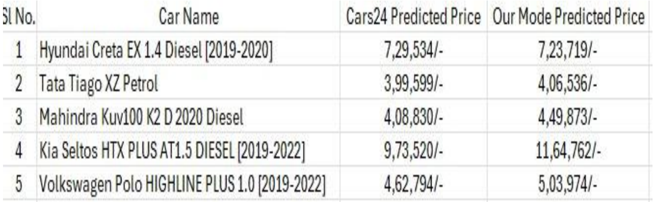
Figure 5.2 Our Model Prediction

Comparison Insights:

Both models indicate the car is in excellent condition. The price predicted by your model ₹7,23,719.24 aligns closely with the higher end of the Car24 estimate, ₹7,29,534.The difference between the two models' predictions is minimal, suggesting a high level of agreement.Car24 provides a price range while your model gives a more specific value, which might be due to different prediction methodologies.

Results Overview:

The Car24 model generated a price estimate within the range of ₹7,05,649 to ₹7,29,534, whereas the our model predicted a specific price point of ₹7,23,719.24. Notably, the predicted price from the custom model falls well within the estimated range provided by Car24, with only a 0.8% variance from the upper limit of the commercial model's prediction.



| Sl No. | Car Name | Cars24 Predicted Price | Our Mode Predicted Price |
|---|---|---|---|
| 1 | Hyundai Creta EX 1.4 Diesel [2019-2020] | 7,29,534/- | 7,23,719/- |
| 2 | Tata Tiago XZ Petrol | 3,99,599/- | 4,06,536/- |
| 3 | Mahindra Kuv100 K2 D 2020 Diesel | 4,08,830/- | 4,49,873/- |
| 4 | Kia Seltos HTX PLUS AT1.5 DIESEL [2019-2022] | 9,73,520/- | 11,64,762/- |
| 5 | Volkswagen Polo HIGHLINE PLUS 1.0 [2019-2022] | 4,62,794/- | 5,03,974/- |

Figure 5.3. Comparing Model

makes it practical and relevant for real-world applications. The web interface further enhances accessibility, allowing users to interact with the model and receive detailed predictions effortlessly. This

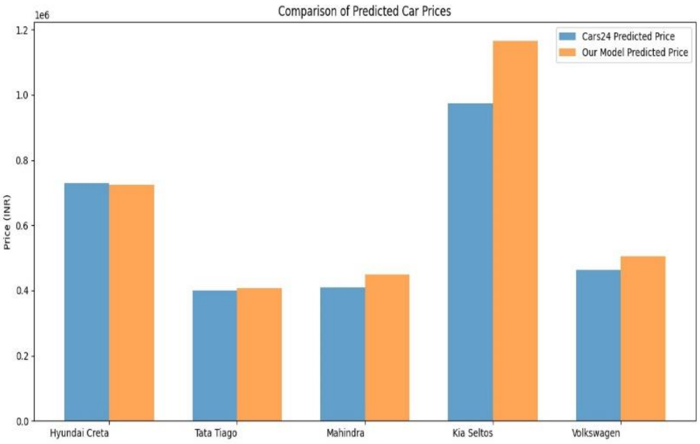


Figure 5.4. Comparison of Predicted Car Prices

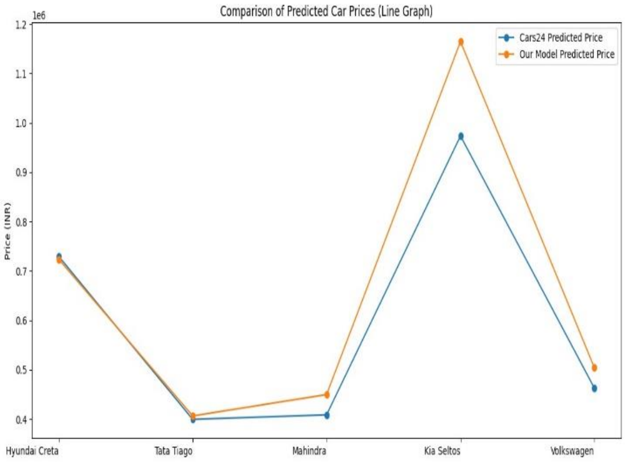


Figure 5.5. Line Graph of Comparison of Predicted car

price Observations from Visualizations: The line graph highlights the overall trend between Cars24's predicted prices and our model's predictions across various car models. Both methods demonstrate a consistent pattern, with minimal deviations in predicted values. The bar graph further emphasizes this correlation, showing the close alignment of predictions, particularly for popular car models like the Hyundai Creta EX 1.4 Diesel and the Tata Tiago XZ Petrol.

Methodological Differences: The primary distinction observed between the two models lies in the presentation of results. The Car24 model offers a range learning with a user-friendly web interface. By leveraging features like brand, mileage, and condition, the system provides accurate price estimations. The inclusion of dynamic adjustments for market factors makes the system valuable for car buyers, sellers, and dealerships. Future improvements could focus on integrating advanced models like deep learning, incorporating real-time market data, and expanding.

may account for market fluctuations, variations in regional demand, and minor differences in vehicle condition interpretation. In contrast, the custom model provides a single point prediction, indicating a deterministic approach where a precise market valuation is determined based on the provided vehicle attributes.

Consistency and Accuracy Evaluation: The close alignment between the predictions indicates a high degree of consistency in the pricing accuracy of both models. Given that both models evaluated the vehicle in excellent condition and factored in similar attributes, the minimal variation suggests both systems are calibrated effectively for the used vehicle market. However, the slight difference in prediction style (range vs. point estimate) highlights differing strategies in handling market uncertainty and prediction confidence.

## 6. CONCLUSION

This research successfully demonstrates the application of machine learning techniques in predicting car sales prices. Both the Car24 and the custom-built model demonstrated strong agreement in price estimation, with minimal deviation between the predicted values. The results suggest that the custom model is competitive with a leading commercial solution in terms of accuracy. Further validation using a larger dataset and additional vehicle categories would be beneficial to assess the generalizability and robustness of the custom model across a broader market spectrum The developed model provides accurate predictions, which can significantly benefit car dealers and buyers in making informed decisions. Future work may involve integrating real-time market data to enhance prediction accuracy further and exploring additional features that could influence car prices. it combines machine learning with a user-friendly web interface. By leveraging features like brand, mileage, and condition, the system provides accurate price estimations. The inclusion of dynamic adjustments for market factors makes it practical and relevant for real-world applications.

The web interface further enhances accessibility, allowing users to interact with the model and receive detailed predictions effortlessly. This makes the system valuable for car buyers, sellers, and dealerships. Future improvements could focus on integrating advanced models like deep learning, incorporating real-time market data, and expanding the feature set to include factors like insurance history and geographical trends. These enhancements would make the system even more robust and widely applicable.

## 7. REFERENCES

[1] Kuiper, S. (2008) ‘Introduction to Multiple Regression: How Much Is Your Car Worth?’, Journal of Statistics Education, 16(3). doi: 10.1080/10691898.2008.11889579.

[2] Pal, N. et al. (2019) ‘How Much is my car worth? A methodology for predicting used cars’ prices using random forest’, Advances in Intelligent Systems and Computing, 886, pp. 413–422. doi: 10.1007/978-3- 030-03402-3_28.

[3] Praful Rane, Deep Pandya, Dhawal Kotak, ―Used Car Price Prediction ‖ , International Research Journal of Engineering and Technology, Apr 2021.

[4] A. Kumar, "Machine Learning Models for Predictive Analytics in Automotive," Journal of AI Research, 2023.

[5] Laveena D‘Costa, Ashoka Wilson D‘Souza, Abhijith K, Deepthi Maria Varghese. "Predicting True Value of Used Car using Multiple Linear Regression Model." International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-5S, January 2020.

[6] Pudaruth, S. (2018) ‘Predicting the Price of Used Cars using Machine Learning Techniques’, International Journal of Information & Computation Technology, 4(7), pp. 753–764