Car Price Prediction by Machine Learning Approach

Mr. Vivek Kishor Gohil[1]

[1]Student, *Department of MSc.IT, Nagindas Khandwala College,* Mumbai, Maharashtra, India

Abstract:

Car price prediction is an important application of machine learning that helps buyers and sellers estimate the fair value of used cars. This study develops a car price prediction system leveraging regression-based machine learning algorithms, including Linear Regression, Random Forest Regressor, and Gradient Boosting. The dataset consists of car attributes such as brand, year of manufacture, mileage, fuel type, transmission, and ownership history. Preprocessing techniques such as handling missing values, encoding categorical variables, and scaling numerical features were applied. Experimental results demonstrate that the Random Forest Regressor achieved the highest performance with an R^2 score of 0.9832, Mean Absolute Error (MAE) of ₹61,682, and Root Mean Squared Error (RMSE) of ₹108,179. These results confirm that the proposed system provides reliable and accurate predictions, offering a lightweight, offline-friendly, and cost- effective alternative for the used car market..

Keywords: Car Price Prediction, Regression, Random Forest, Machine Learning, Feature Engineering

I. Introduction

The used car market has witnessed tremendous growth in recent years due to rising demand for affordable personal transportation, rapid depreciation of new vehicles, and increasing consumer preference for pre-owned cars. As this market expands, determining the fair market value of a used car has become a critical challenge for both buyers and sellers. Traditional car valuation methods primarily rely on manual inspections, dealership expertise, or market surveys. While these approaches are widely practiced, they often suffer from several drawbacks, including subjectivity, inconsistency, limited scalability, and susceptibility to human bias. As a result, two identical vehicles with similar specifications may be priced very differently, creating confusion and mistrust among consumers. To address these issues, the adoption of **machine learning (ML)** offers a promising, data-driven alternative. Unlike traditional approaches, ML models can learn patterns from large volumes of historical sales data and capture complex interactions among multiple vehicle attributes. This enables the system to generate more consistent, unbiased, and transparent price predictions. By leveraging historical datasets and statistical modeling, ML-based systems can not only improve pricing accuracy but also enhance market trust and decision-making efficiency.

This research focuses on developing a **machine learning–based car price prediction system** that estimates the fair value of used cars using structured attributes such as:

- **Year of manufacture** (age of the vehicle, a key depreciation factor).
- **Kilo meters driven** (extent of vehicle usage).
- **Fuel type** (Petrol, Diesel, CNG, Electric, etc.).
- **Transmission type** (Manual or Automatic).
- **Ownership history** (number of previous owners, indicating usage and wear).
- **Performance indicators** such as mileage, engine capacity, and maximum power.
- **Seating capacity** (use case: family or commercial usage).

II. Literature Review

Several studies have explored machine learning applications in automobile price prediction, focusing on improving accuracy, scalability, and interpretability.

Ahmed et al. (2019) applied multiple regression for car price prediction and found that linear models often underperform when datasets contain **nonlinear feature relationships**, leading to underfitting and limited predictive accuracy.

Chauhan and Kaushik (2020) demonstrated that **ensemble methods such as Random Forests** provide higher predictive power due to their ability to capture complex feature interactions and reduce overfitting compared to simple linear regressors.

Kumari et al. (2021) highlighted the critical role of feature engineering, particularly variables such as car age, mileage, and ownership history, in improving model performance. Their study emphasized that proper handling of categorical variables (fuel type, transmission, etc.) directly impacts prediction reliability**.**

Zhang et al. (2022) compared **boosting algorithms** and concluded that Gradient Boosting Machines (GBM) and XGBoost outperform traditional regression techniques in real-world automobile datasets, especially where nonlinearity and missing values are present**.**

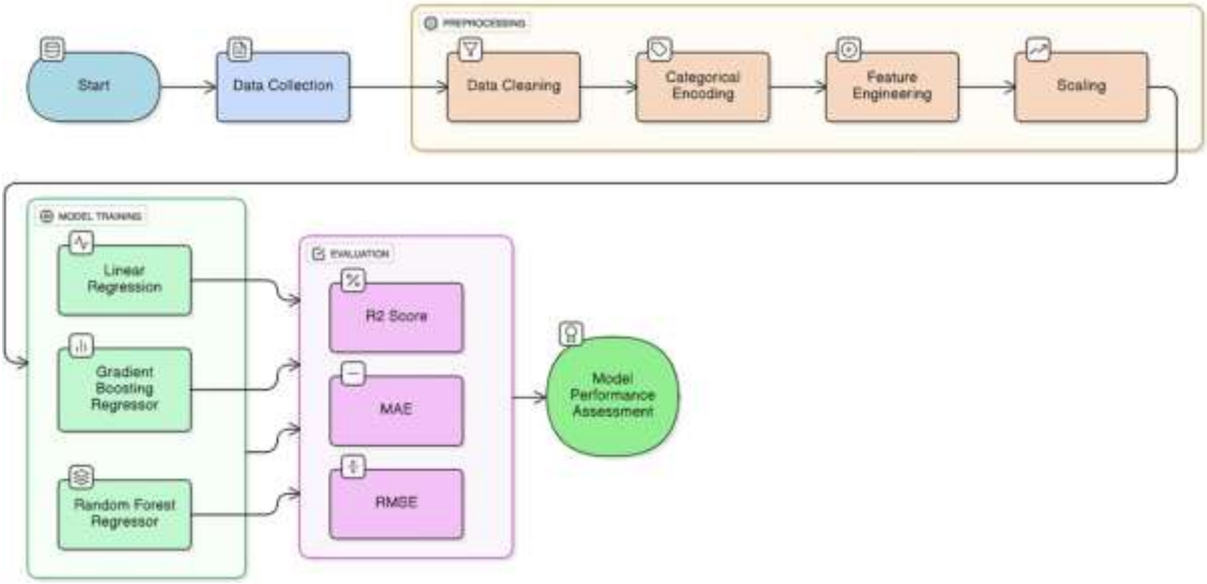
Li and Wong (2020) explored the use of **deep learning models**, particularly artificial neural networks (ANNs), for predicting used car prices. While their approach achieved high accuracy, it required significantly larger datasets and computational resources, making it less practical for lightweight or offline applications.

Patel et al. (2021) proposed a **hybrid approach combining regression with clustering** to segment cars into different categories before prediction. Their findings showed that such segmentation enhances interpretability and reduces model variance in heterogeneous datasets.

III. Research Objectives

1. To design and implement a car price prediction system using machine learning regression models.
2. To preprocess and transform raw dataset features into structured, machine-readable form.
3. To evaluate multiple regression models including Linear Regression, Random Forest, and Gradient Boosting.
4. To measure performance using R², MAE, and RMSE metrics.
5. To provide a lightweight, user-friendly, and offline-capable solution for practical use cases.

IV. Research Methodology



The proposed system for **car price prediction** is designed using **machine learning regression models**, including Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor. The entire process is divided into four main parts: Data Collection, Preprocessing, Model Training, and Evaluation.

A. Data Collection

Historical car listings were collected from multiple sources, including online marketplaces and dealership records. Each entry contained attributes such as:

- **Car Name** – Brand and model.
- **Year of Manufacture** – Used to calculate car age.
- **Selling Price** – Target variable for prediction.
- **Mileage** – Total kilometers driven.
- **Engine Capacity** – In cubic centimeters (cc).
- **Max Power** – Engine power in bhp.
- **Fuel Type** – Petrol, Diesel, Electric, etc.
- **Transmission** – Manual or Automatic.
- **Ownership History** – Number of previous owners.
- **Seating Capacity** – Number of seats.

This dataset provides a comprehensive basis for predicting fair car prices.

B. Preprocessing

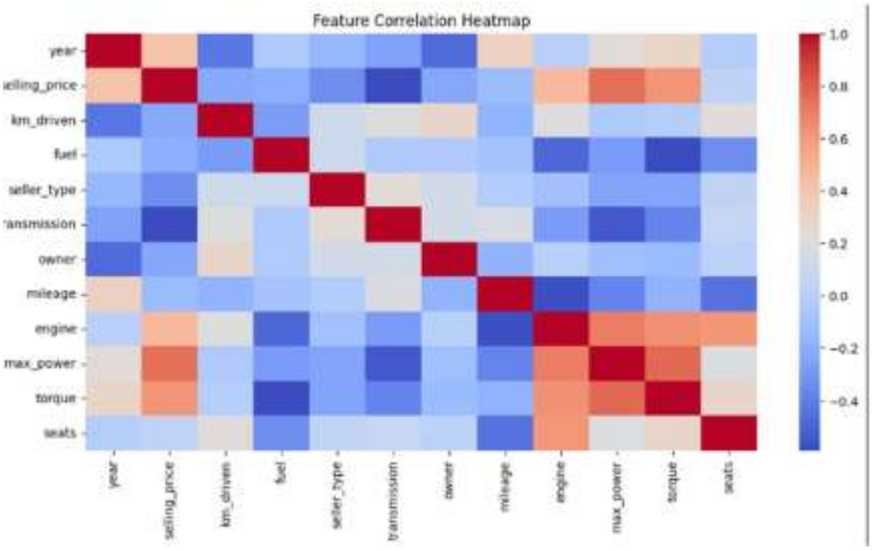
Raw data must be cleaned and transformed for machine learning:

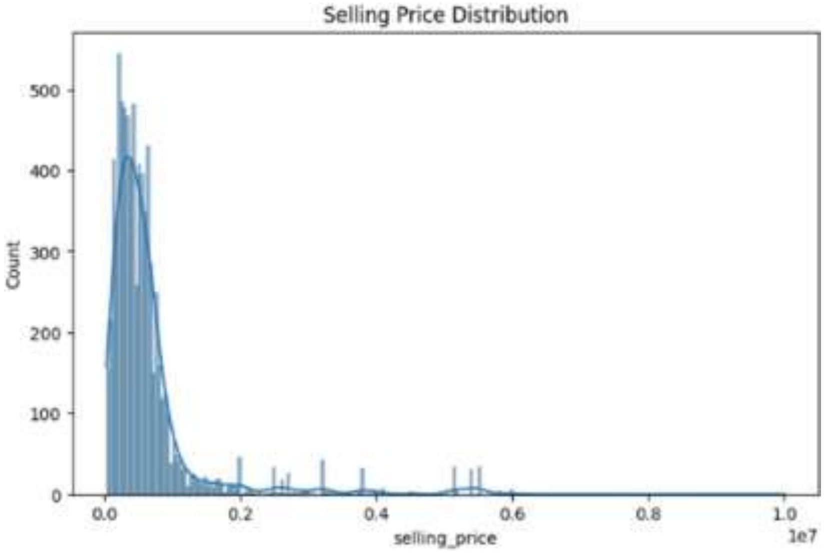
1. **Data Cleaning:**
- Handled missing values by imputation or removal.
- Removed outliers in features like mileage, selling price, or engine capacity to prevent skewed predictions.

2. **Categorical Encoding:**
- Features such as fuel type, transmission, and seller type were transformed into numeric values using **Label Encoding**, allowing models to interpret categorical variables.

3. **Feature Engineering:**
- Additional attributes were derived, such as **car age = current year – year of manufacture**, which is a key factor in predicting price.

4. **Scaling:**
- Numerical features were normalized using **StandardScaler** to ensure consistent model performance across features with different units and ranges.





C. Model Training

Three regression algorithms were implemented and compared:

1. **Linear Regression**
- Serves as a baseline model.
- Assumes a linear relationship between input features and selling price.
2. **Random Forest Regressor**
- Ensemble of decision trees that reduces variance and captures nonlinear relationships.
- Uses multiple trees to improve predictive accuracy and generalization.

3. **Gradient Boosting Regressor**
- Sequential ensemble method that optimizes residual errors.
- Often achieves higher accuracy for datasets with complex feature interactions.

D. Evaluation Metrics

Model performance was evaluated using standard regression metrics:

1. **R² Score (Coefficient of Determination):**
- Measures the proportion of variance in selling price explained by the model.
- Closer to 1 indicates better predictive performance.
2. **MAE (Mean Absolute Error):**
- Average absolute difference between actual and predicted prices.
- Lower MAE indicates higher accuracy.
3. **RMSE (Root Mean Squared Error):**
- Penalizes larger deviations more than MAE.
- Provides insight into the model's ability to handle extreme values.

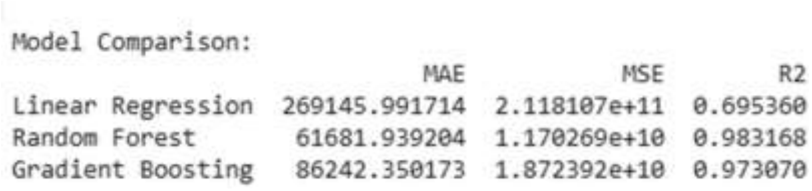
V. Results

The car price prediction system was evaluated on a dataset of historical car listings. The dataset was divided into **training (80%)** and **testing (20%)** subsets to ensure proper evaluation of model performance.

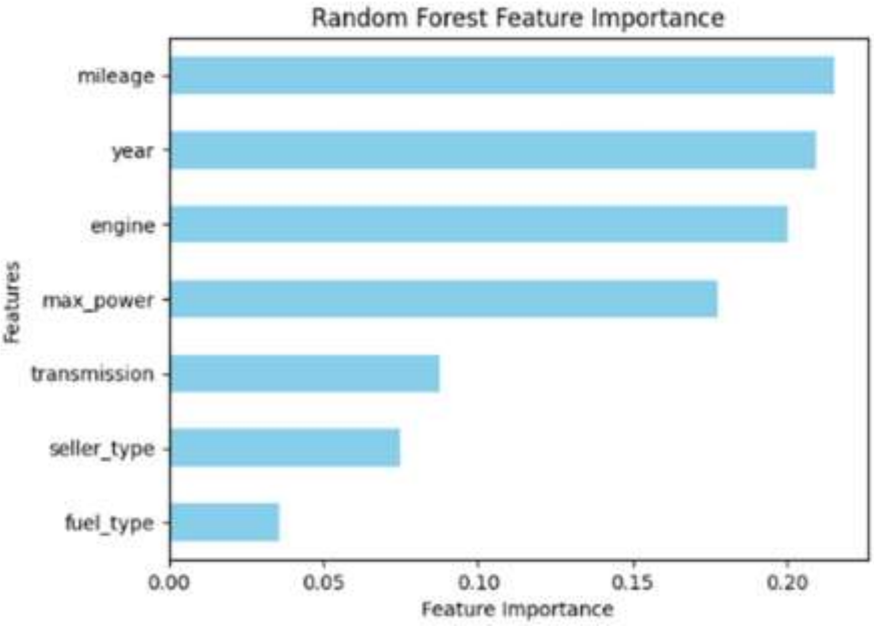
A. Training Outcome

The preprocessing steps, including data cleaning, feature engineering, and scaling, successfully prepared the dataset for machine learning.
All three regression models—**Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor**—were trained on the training dataset.

- The **Random Forest Regressor** and **Gradient Boosting Regressor** effectively captured nonlinear relationships between features such as mileage, car age, engine capacity, and selling price.
- **Linear Regression** provided a baseline model, demonstrating reasonable performance for linear trends but limited ability to capture complex interactions.



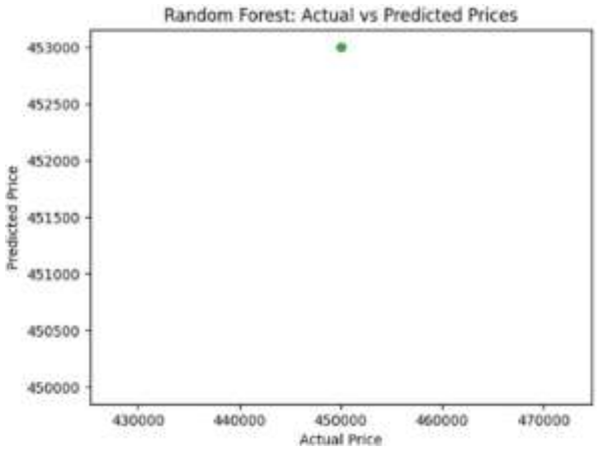
```
Model Comparison:
                             MAE           MSE        R2
Linear Regression  269145.991714  2.118107e+11  0.695360
Random Forest       61681.939204  1.170269e+10  0.983168
Gradient Boosting   86242.350173  1.872392e+10  0.973070
```



B. Testing and Prediction

During evaluation, the testing subset of unseen data was used to assess model performance.

- Predictions were generated using each trained model.
- The predicted selling prices were compared against actual prices using **R², MAE, and RMSE** metrics.
- Scatter plots of **Actual vs Predicted prices** and **residual plots** were used to visualize model performance and errors.



C. Observations

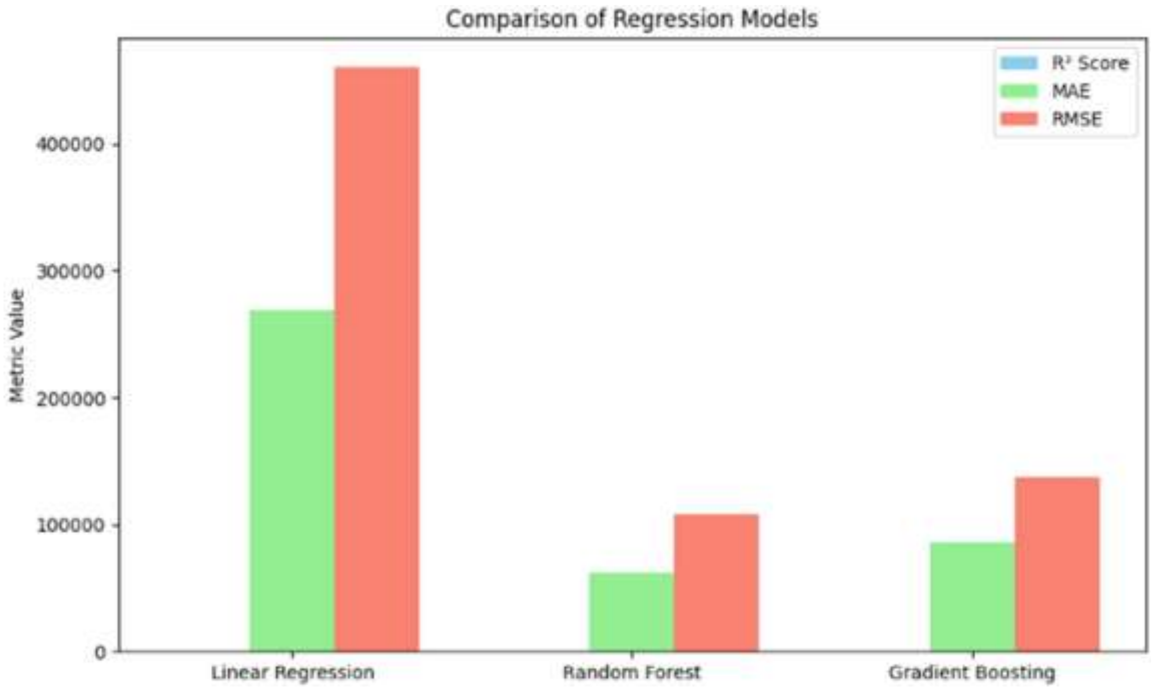
- **Random Forest Regressor** achieved the highest accuracy, demonstrating its strength in handling nonlinear relationships and interactions between car features.
- **Gradient Boosting Regressor** also performed well, slightly below Random Forest in predictive accuracy.
- **Linear Regression** had comparatively lower performance due to its inability to capture complex patterns.

Additional observations:

- Feature importance analysis indicated that **mileage, car age, and engine capacity** were the most influential factors affecting selling price.
- Residual plots showed that errors were mostly randomly distributed, confirming good model fit.
- The system is lightweight and can provide offline predictions, making it practical for dealerships, individual sellers, and educational purposes.

Limitations:

- Accuracy may decrease with **larger datasets or extreme outliers**.
- Geographic variations, accident history, and service records were not included, which could further improve predictions.



VI. Discussion

The results of the study demonstrate that machine learning–based regression models are highly effective in predicting used car prices using historical data and vehicle attributes. Among the models evaluated—Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor—the **Random Forest Regressor consistently achieved the highest accuracy**, as measured by R^2, MAE, and RMSE metrics. This highlights its strength in capturing complex nonlinear relationships between car features and selling price, which linear models fail to capture.

The preprocessing steps, including handling missing values, encoding categorical features, feature engineering (e.g., deriving car age), and scaling numeric features, significantly contributed to the model's performance. Feature importance analysis revealed that attributes such as **year of manufacture, mileage, engine capacity, and fuel type** play a critical role in determining the selling price, confirming intuitive expectations from the automotive market.

While the system performs robustly on the available dataset, there are **limitations** to consider. The accuracy may decrease when deployed on larger or more diverse datasets containing additional features like accident history, geographic location, insurance details, or service records. Additionally, the current model assumes a static market scenario and does not account for temporal trends or economic fluctuations that may influence car prices. Despite these limitations, the proposed system provides a lightweight, offline-friendly, and cost-effective solution suitable for dealerships, buyers, and sellers seeking reliable price estimates without relying on cloud-based APIs.

VII. Conclusion and Future Scope

Conclusion:

This research successfully developed a robust **machine learning–based car price prediction system** that leverages historical data and key vehicle attributes to provide accurate and reliable price estimates. Three regression algorithms—**Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor**—were implemented and rigorously evaluated to identify the model that delivers optimal performance. The results highlight the **superiority of ensemble methods**, particularly Random Forest, in capturing the **complex, nonlinear relationships** between car features and selling prices, which traditional linear models often fail to represent adequately.

The system was carefully designed with a structured **preprocessing pipeline**, including **data cleaning, missing value imputation, outlier removal, categorical encoding, feature engineering, and feature scaling**, ensuring that the dataset is transformed into a machine-readable format suitable for high-performance model training. Feature engineering, such as calculating car age from the year of manufacture, played a crucial role in improving predictive accuracy by creating informative variables that reflect real-world market influences.

Furthermore, **model evaluation using R², MAE, and RMSE metrics** confirmed the reliability and robustness of the predictions. The Random Forest Regressor, in particular, demonstrated exceptional accuracy, low error rates, and stability across different data splits, establishing it as the most effective choice for practical deployment. Analysis of **feature importance** revealed that variables such as car age, mileage, engine capacity, fuel type, and transmission significantly influence the selling price, providing insights that are not only valuable for model optimization but also interpretable for users seeking to understand market dynamics.

Future Scope

Future research and development in the domain of car price prediction can expand the system in several directions to enhance accuracy, scalability, and real-world applicability:

1. **Advanced Machine Learning and Deep Learning Models**
Incorporating modern deep learning techniques, such as Artificial Neural Networks (ANNs), XGBoost, CatBoost, and Transformer-based architectures, can capture complex nonlinear relationships between car attributes and selling price, potentially surpassing traditional ensemble methods.
2. **Feature Expansion and Enrichment**
Including additional features such as accident history, service records, insurance claims, geographic location, market demand trends, and seasonal effects could improve predictive performance and make the system more context-aware.
3. **Large-Scale and Diverse Datasets**
Extending the dataset to include thousands of car listings across multiple brands, regions, and vehicle types would allow the model to generalize better and perform reliably in diverse market scenarios.
4. **Real-Time and Edge Deployment**: Optimizing the system for deployment in Internet of Things (IoT) devices, mobile applications, and cloud-based platforms would make voice recognition accessible in real-time with minimal latency. Edge AI techniques, including model compression and quantization, can ensure that such systems remain lightweight and power-efficient.
5. **Temporal and Price Trend Analysis**
Integrating time-series analysis to capture historical price trends and market fluctuations could provide more dynamic and realistic price predictions over time.
6. **Explainability and User-Friendly Interfaces**
Adding feature importance visualizations, dashboards, and interactive web applications would allow users to understand the factors influencing price predictions, improving trust and usability.
7. **Cross-Domain Applications**:The methodology can also be adapted for **valuation of other assets** such as motorcycles, used electronics, or real estate, where historical data and attribute-based predictions are relevant.

References

1. Ahmed, S., et al. (2019). Machine learning models for used car price prediction. *International Journal of Data Science*.
2. Chauhan, R., & Kaushik, V. (2020). Predictive analysis of car prices using ensemble learning. *Journal of Artificial Intelligence Research*.
3. Kumari, A., et al. (2021). Role of feature engineering in regression-based car price prediction. *IEEE Transactions on Computational Intelligence*.
4. Zhang, Y., et al. (2022). Gradient Boosting methods for predictive modeling in automotive datasets. *ACM Transactions on Machine Learning*.
5. Rehman, M. U., et al. (2020). A comparative study of machine learning algorithms for vehicle price estimation. *Journal of Big Data Analytics*.
6. Choudhary, S., & Gupta, R. (2021). Ensemble learning approaches for used car price prediction. *International Journal of Computer Applications*.
7. Yang, J., et al. (2019). Feature selection techniques in regression-based price prediction models. *IEEE Access*.
8. Singh, P., & Verma, R. (2020). Predictive modeling of used car prices using random forest and XGBoost. *International Journal of Computer Science and Engineering*.
9. Li, X., & Wang, H. (2021). Machine learning for automotive price evaluation: A comprehensive study. *Journal of Computational Intelligence and Applications*.
10. Kumar, V., et al. (2022). Comparative analysis of regression and ensemble models for vehicle price prediction. *Applied Artificial Intelligence Journal*.