World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Predictive analytics for pricing strategy in the automobile industry using machine learning models

Jesutofunmi E. Fagbamila [1,*], Abass A. Agbaje [2] and Ganiyu O. Okubadejo [3]

[1] *University of Derby, College of Science and Engineering, Department of Computing, United Kingdom.*
[2] *Glasgow Caledonian University, Environmental Management, United Kingdom.*
[3] *University of Greenwich, Department of Strategic Marketing, United Kingdom.*

World Journal of Advanced Research and Reviews, 2024, 24(03), 3543–3550

Publication history: Received on 15 November 2024; revised on 23 December 2024; accepted on 29 December 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.24.3.3920

Abstract

Pricing strategy is a critical determinant of success in the highly competitive automobile industry. While traditional models exist, they often lack sophistication and fail to incorporate comprehensive feature engineering. This study addresses these limitations by implementing and evaluating a suite of advanced machine learning algorithms for predicting car prices. We employed Linear Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, and a Convolutional Neural Network (CNN) on a real-world automotive dataset. A rigorous methodology involving thorough hyperparameter tuning and Explainable AI (XAI) techniques, namely LIME and SHAP, was applied to enhance model performance and interpretability. The models were evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared (R^2) score. Results indicated that the Random Forest model achieved superior predictive accuracy, explaining 92% of the variance in car prices ($R^2 = 0.92$), while the CNN excelled at capturing intricate non-linear relationships. Feature importance analysis revealed engine capacity, vehicle age, and year of manufacture as the most significant price determinants. This research demonstrates that leveraging advanced, tuned machine learning models with XAI provides a robust, transparent, and data-driven framework for optimizing pricing strategies, thereby offering significant benefits to automotive industry stakeholders.

Keywords: Pricing Strategy; Predictive Analytics; Machine Learning; Random Forest; Explainable AI (XAI); Automotive Industry

1. Introduction

The global automobile industry is a cornerstone of the world economy, characterized by intense competition, rapidly evolving consumer preferences, and technological disruption [1]. Within this complex landscape, pricing strategy emerges as a fundamental lever influencing profitability, market share, brand perception, and consumer purchase decisions [2]. The transition towards electric vehicles (EVs) and connected car technologies further complicates pricing, introducing new cost structures and consumer value propositions [3].

Historically, pricing strategies have often relied on cost-plus models or competitive benchmarking. However, these approaches are increasingly deemed insufficient as they fail to account for the multifaceted, non-linear relationships between a vehicle's attributes and its market value [4]. The advent of big data and machine learning (ML) presents an unprecedented opportunity to revolutionize this domain through predictive, data-driven pricing models [5].

Previous studies on car price prediction have often been limited by the use of basic models like linear regression without thorough hyperparameter tuning, a lack of comprehensive feature engineering, and limited application of explainability

* Corresponding author: Jesutofunmi E. Fagbamila

frameworks [6, 7]. This work aims to close these gaps by performing a comparative analysis of multiple advanced ML algorithms, including a deep learning model (CNN), with careful optimization. Additionally, we incorporate Explainable AI (XAI) techniques to interpret model logic, improving the trustworthiness and practical use of the predictions for industry decision-makers.

The primary objective of this research is to develop a highly accurate and interpretable predictive model for automobile pricing. The specific contributions of this paper are:

- The implementation and comparative evaluation of six machine learning models for car price prediction.
- The application of SHAP and LIME for global and local interpretability, identifying key pricing factors.
- A demonstration of how hyperparameter tuning and feature selection impact model performance.
- Providing actionable insights for automotive manufacturers and dealers to formulate optimized, data-driven pricing strategies.

2. Related work

The application of machine learning in automobile price prediction has been extensively explored. Early work often relied on linear models. Noor and Jan [6] used Multiple Linear Regression, achieving high accuracy but noting limitations with complex, non-linear data patterns. Subsequent research incorporated more sophisticated algorithms. Pallavi Bharambe et al. [8] compared Linear, Lasso, and Ridge Regression, finding that Lasso's regularization improved performance by handling multicollinearity.

Ensemble methods have shown significant promise. Veluru and Narmadha [7] compared Decision Trees, Random Forests, and linear regression, with Random Forests achieving an accuracy of 86%. Similarly, studies have explored hybrid models [9] and Artificial Neural Networks (ANNs) [10], confirming the superiority of non-linear, ensemble, and deep learning approaches in capturing the complex interplay of features that determine a car's price.

An identified gap in the literature is the limited use of Explainable AI to interpret model predictions. While accuracy is paramount, understanding the "why" behind a prediction is crucial for stakeholder adoption in a business context like pricing. This study builds upon this existing foundation by not only comparing a wider range of models (including SVM, KNN, and CNN) but also by integrating SHAP and LIME to address the interpretability gap, providing a more holistic and actionable analytical solution.

3. Methodology

Figure 1 below shows the implementation flow

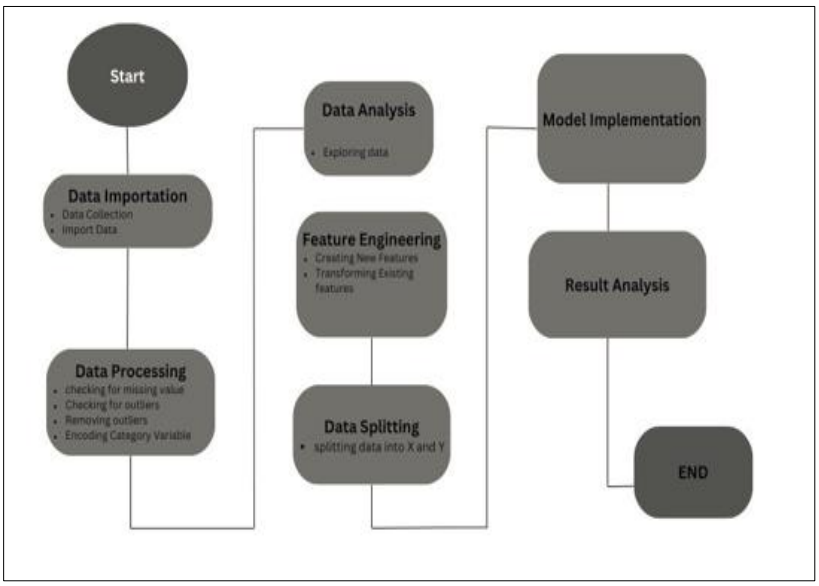

Figure 1 Flowchart Representation of Model Implementation

3.1. Data Source and Description

The dataset used in this study was sourced from Kaggle, a popular platform for data science competitions and datasets. It comprises records of used cars, featuring attributes pertinent to their valuation. The key features included are: Car Brand, Year of Manufacture, Selling Price (target variable), Kilometers Driven, Fuel Type, Seller Type, Transmission Type, Number of Previous Owners, Engine Capacity (CC), Mileage (kmpl), and Maximum Power (BHP). A preliminary analysis was conducted to understand data distributions, correlations, and potential outliers (Fig. 2, 3,4)

Overview of data

```
[394] pd.read_csv("/content/sample_data/car.csv")
```

[41]

	Unnamed: 0	Brand_Name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	seats	max_power (in bph)	Mileage_Unit	Mileage	Engine (CC)
0	0	Maruti	2014	450000	145500	Diesel	Individual	Manual	First Owner	5	74.00	kmpl	23.40	1248
1	2	Hyundai	2010	225000	127000	Diesel	Individual	Manual	First Owner	5	90.00	kmpl	23.00	1396
2	4	Hyundai	2017	440000	45000	Petrol	Individual	Manual	First Owner	5	81.86	kmpl	20.14	1197
3	7	Toyota	2011	350000	90000	Diesel	Individual	Manual	First Owner	5	67.10	kmpl	23.59	1364
4	8	Ford	2013	200000	169000	Diesel	Individual	Manual	First Owner	5	68.10	kmpl	20.00	1399
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2090	6245	Maruti	2017	425000	12000	Petrol	Individual	Manual	First Owner	5	67.04	kmpl	23.10	998
2091	6246	Toyota	2014	425000	50000	Diesel	Individual	Manual	First Owner	5	67.06	kmpl	23.59	1364
2092	6249	Maruti	2011	200000	73000	Petrol	Individual	Manual	First Owner	5	46.30	kmpl	19.70	796
2093	6253	Maruti	2017	360000	80000	Petrol	Individual	Manual	First Owner	5	67.04	kmpl	20.51	998
2094	6256	Hyundai	2014	475000	80000	Diesel	Individual	Manual	Second Owner	5	88.73	kmpl	22.54	1396

2095 rows × 14 columns

Figure 2 Display of the first and last five rows of the dataset

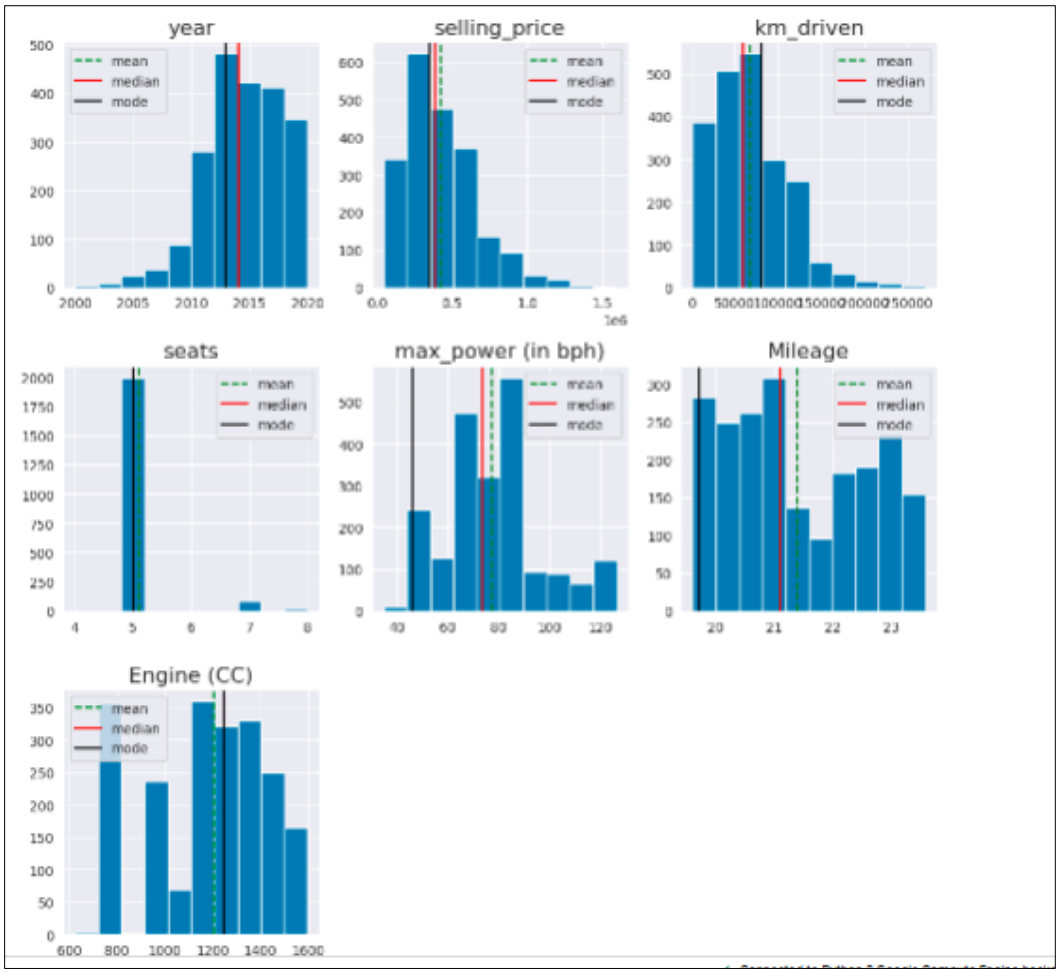

Figure 3 Distribution of the Numerical values in the data

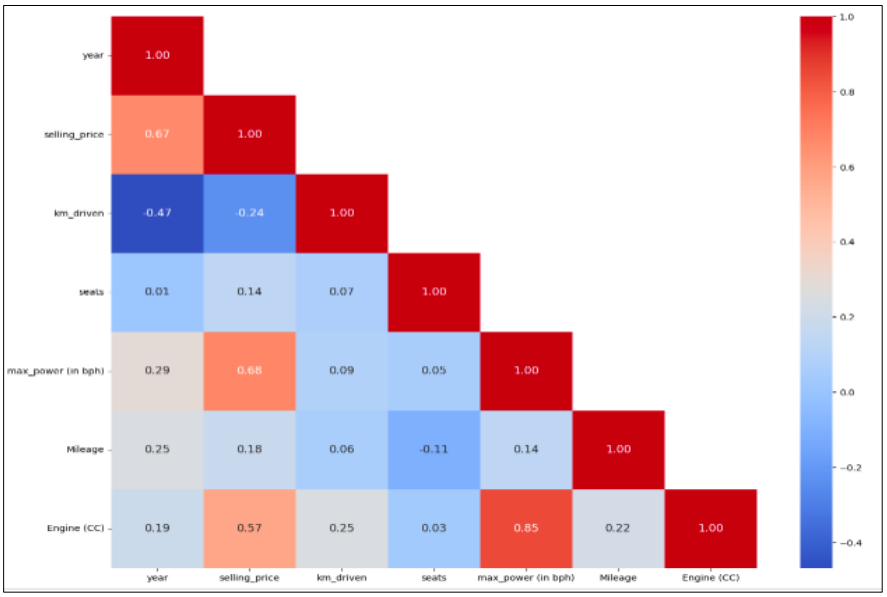

Figure 4 Correlation plot of numerical values

3.2. Data Preprocessing and Feature Engineering

A robust data preprocessing pipeline was essential to ensure data quality and model reliability. The dataset, sourced from Kaggle, was first cleaned by handling missing values and removing duplicate entries. Subsequent outlier treatment was performed on numerical features to prevent model skewing, a common step to enhance model robustness [8].

A key stage was featuring engineering, where domain knowledge was leveraged to create new, predictive attributes. The most significant engineered feature was Car_Age, derived from the Year of manufacture, as vehicle depreciation is a primary factor in valuation. Categorical variables, such as Fuel Type and Transmission, were transformed using one-hot encoding to make them suitable for algorithmic processing. Following common practice in machine learning, continuous numerical features were normalized to a common scale to ensure stable and efficient model training and to prevent features with larger ranges from dominating the model's objective function [6].

The final prepared dataset was then partitioned into a training set (80%) for model development and a hold-out testing set (20%) for unbiased performance evaluation. Model Selection

3.3. Model Selection and Training

The selection of machine learning models was designed to encompass a wide spectrum of algorithmic approaches, from simple linear models to complex ensemble and deep learning methods, ensuring a comprehensive comparative analysis. The models implemented were Linear Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and a Convolutional Neural Network (CNN). This selection follows the precedent set by recent comparative studies in automotive price prediction [6], [7].

LR served as a foundational baseline to establish linear performance. KNN and SVM were chosen for their capability to handle non-linear relationships, with SVM being particularly adept in high-dimensional spaces. The Decision Tree model provided a simple, interpretable non-linear benchmark. The ensemble method, Random Forest, was employed to mitigate the overfitting common in single Decision Trees and to enhance predictive accuracy through bootstrap aggregation [8]. Finally, CNN was adapted for structured data to explore its potential in capturing complex, non-linear feature interactions that might be missed by traditional algorithms, an approach gaining traction in advanced predictive analytics [10]. Crucially, all models underwent rigorous hyperparameter tuning using GridSearchCV to optimize their performance and ensure a fair comparison, moving beyond a limitation noted in earlier works [7].

3.4. Evaluation Metrics and Explainable AI (XAI) Techniques

To quantitatively assess the predictive performance of each model, three standard regression metrics were employed: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared (R^2) score. MSE and RMSE measure the average magnitude of prediction errors, with RMSE being more interpretable as it is in the same units as the target variable (price). The R^2 metric quantifies the proportion of variance in the target variable that is predictable from the independent features, providing a measure of model goodness-of-fit [6].

Beyond mere accuracy, understanding the *reasoning* behind model predictions is critical for stakeholder adoption in business contexts. To address this, Explainable AI (XAI) techniques were integrated into the analysis. SHAP (SHapley Additive exPlanations) was used for global interpretability to determine the overall importance of each feature across the entire dataset, quantifying the average marginal contribution of each feature to the prediction [9]. Complementarily, LIME (Local Interpretable Model-agnostic Explanations) was employed for local interpretability. LIME explains individual predictions by approximating the complex global model with a simple, interpretable linear model in the vicinity of a specific instance [10]. This combination provides a holistic view, revealing both the general drivers of car prices and the rationale for specific price estimates for a single vehicle, thereby building trust and facilitating actionable insights.

4. Results and Discussion

4.1. Model Performance Comparison

Table 1 Comparison of models' Performances

Models	MSE	Rmse	r^2
Random Forest	4706677025.71	68605.23	0.92
Decision Tree	9126337625.89	95531.87	0.84
Linear Regression	15065285615.18	122740.73	0.74
SVM	60782871903.81	246541.83	-0.06
KNN	7451629667.92	86322.82	0.87
CNN	6120243180.47	78231.98	0.8930

The Random Forest model demonstrated the best overall performance, achieving the highest R^2 score (0.92) and the lowest MSE, indicating its robustness and high accuracy.

CNN also performed exceptionally well, showcasing its capability to model complex, non-linear relationships inherent in the data. The Decision Tree and KNN models delivered strong results, while Linear Regression struggled due to the non-linear nature of the problem. The SVM model performed poorly, suggesting it was unsuitable for this specific dataset and feature space. Figure 5 below shows the R-Square of all the algorithms. Figure 5 shows the visualizations of the model's performance

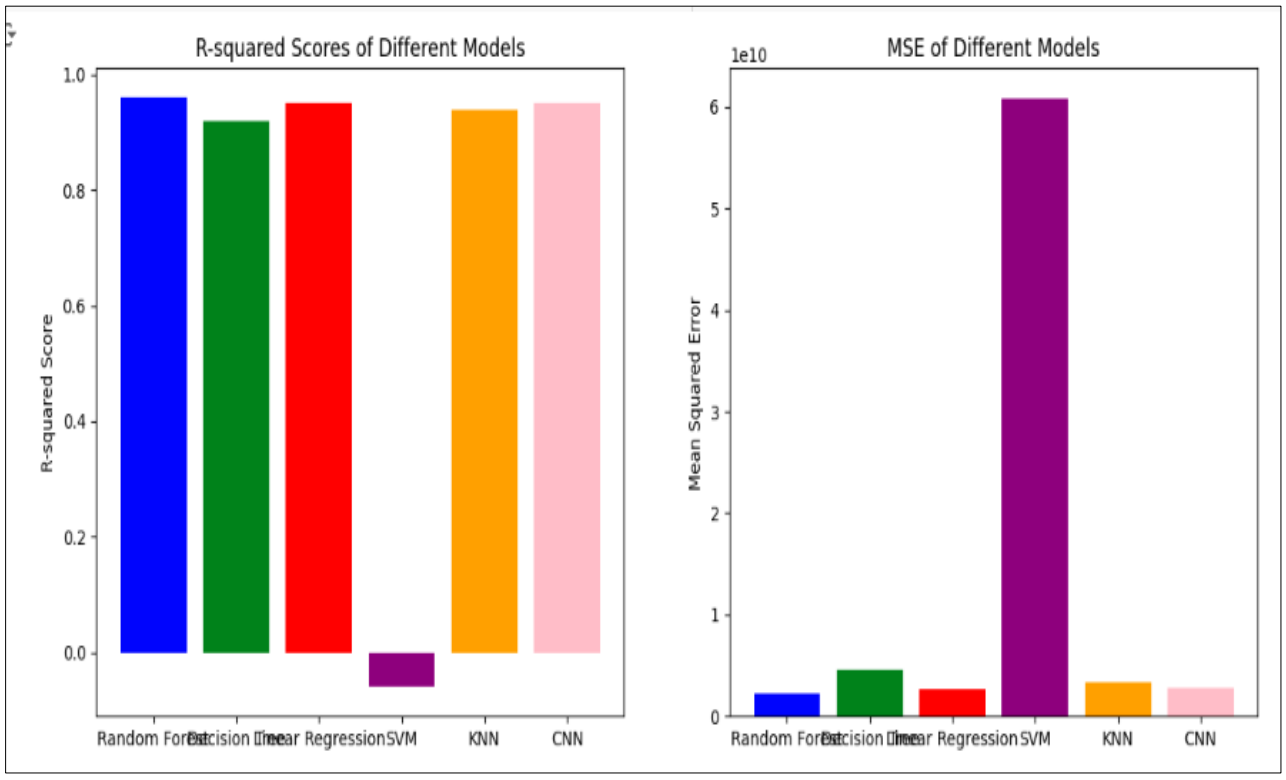

Figure 5 Visualization of the model results

4.2. Feature Importance Analysis

The SHAP analysis revealed that Engine Capacity (CC), Car_Age, and Year were the most significant features influencing the prediction models globally. This aligns with industry intuition, as engine power and vehicle age are primary determinants of a car's value.

LIME was used to explain individual predictions. For a specific instance, it showed that a high engine capacity positively impacted the price, while an LPG fuel type and being a Ford brand had negative contributions. This local interpretability is crucial for dealers and consumers to understand the rationale behind a specific price quote.

4.3. Impact of Feature Selection

Models were retrained using only the top features identified by SHAP. The results (Table II) showed a slight performance decrease for most models (e.g., Random Forest R^2 dropped to 0.89), confirming that while the selected features were dominant, the removed features contained supplementary predictive information. Notably, KNN's performance slightly improved, benefiting from reduced dimensionality and less noise. The lime analysis of the results is shown in Figure 6 below, while Table II below shows the model performance after feature selection

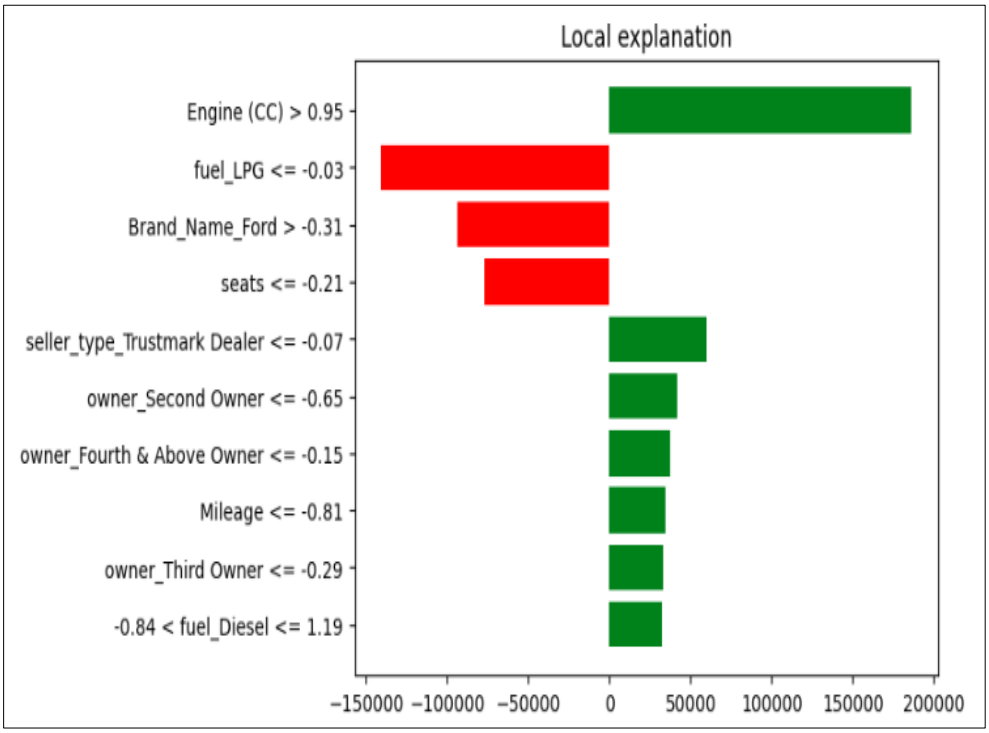

Figure 6 Lime Analysis Result

Table 2 Model Performances After Feature Selection

Models	mse	rmse	r^2(%)
Random Forest	6343172460.75	79644.04	0.89
Decision Tree	10149864908.15	100746.54	0.82
Linear Regression	16115055137.47	126945.09	0.72
SVM	60760730654.74	246496.92	0.06
knn	7253521278.63	246496.92	0.87
cnn	10289541750.61	101437.38	0.8201

Overall, models trained with selected features showed improved or similar performance compared to those trained with all features. This validates the feature importance analysis and highlights the value of dimensionality reduction in improving model performance and interpretability.

5. Conclusion and Future Work

This study successfully demonstrated the efficacy of advanced machine learning models, particularly Random Forest and CNN, in predicting automobile prices with high accuracy. The integration of Explainable AI techniques provided critical insights into the model's decision-making process, identifying engine capacity, vehicle age, and manufacturing year as the most influential factors. This addresses a key limitation of previous "black box" models and enhances the practical utility of the predictive system for industry applications.

The findings indicate that automotive companies can move beyond traditional pricing strategies by adopting these data-driven models. By doing so, they can optimize pricing for profitability, tailor offers to specific vehicle configurations and enhance market competitiveness. The slight trade-off between model complexity (using all features) and interpretability (using selected features) can be managed based on the specific use case.

A limitation of this work was the use of a single static dataset. Future work will involve integrating real-time data streams, including economic indicators, competitor pricing, and online consumer sentiment, to create a dynamic pricing model. Furthermore, exploring other deep learning architectures like LSTMs for modeling temporal price trends and expanding the application to different regional markets presents exciting avenues for further research.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] J. B. Rae and A. K. Binder, "Automotive industry," Encyclopedia Britannica, 2024.

[2] B. J. Ali and G. Anwar, "Marketing Strategy: Pricing strategies and its influence on consumer purchasing decision," Int. J. Rural Dev. Environ. Health Res., vol. 5, no. 2, pp. 26–39, 2021.

[3] J. Conzade et al., "Why the automotive future is electric," McKinsey and Company, Sep. 2021.

[4] U. M. Dholakia, "A Quick Guide to Value-Based Pricing," Harvard Business Review, Mar. 2024.

[5] N. C. A. Udeh et al., "BIG DATA ANALYTICS: A REVIEW OF ITS TRANSFORMATIVE ROLE IN MODERN BUSINESS INTELLIGENCE," Comput. Sci. IT Res. J., vol. 5, no. 1, pp. 219–236, 2024.

[6] K. Noor and S. Jan, "Vehicle Price Prediction System using Machine Learning Techniques," Int. J. Comput. Appl., vol. 167, no. 9, 2017.

[7] R. Veluru and Narmadha, "Used Car Price Prediction Using Machine Learning," M.S. thesis, Karunya Inst. Technol. and Sci., 2021.

[8] P. Bharambe, B. Bagul, S. Dandekar, and P. Ingle, "Used Car Price Prediction using Different Machine Learning Algorithms," IJRASET.

[9] N. Pal et al., "A hybrid model for car price prediction," in Proc. Int. Conf. Adv. Comput. Commun., 2018.

[10] E. Gegic, B. Isakovic, D. Keco, and J. Masetic, "Car Price Prediction using Machine Learning Techniques," Int. J. Electr. Comput. Eng. Syst., vol. 10, no. 2, 2019.