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 11, November 2025

DOI: 10.17148/IJARCCE.2025.141138

A Machine Learning Framework for Automotive Price Prediction and Revenue Forecasting

Arjun Kaymala[1], R Divya[2], Tamizhselvan S.P[3], G. Paavai Anand[4]

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India[1,2,3,4]

Abstract: The automotive market's increasing digitalization necessitates accurate, data-driven vehicle valuation models to en- hance market transparency and support strategic decision-making. This paper presents a machine learning framework designed to predict car prices based on a comprehensive set of technical and physical specifications. The core of this research is a comparative analysis of a baseline Linear Regression model against a more sophisticated Random Forest Regressor to evaluate their predictive efficacy. Using a structured dataset of over 200 vehicle records, our methodology incorporates a robust preprocessing pipeline, including one-hot encoding for categorical features and standard scaling for numerical attributes. The empirical results demonstrate the superior performance of the Random Forest model, which achieved a coefficient of determination (R2) of 0.96, alongside a Root Mean Squared Error (RMSE) of 1791.80 and a Mean Absolute Error (MAE) of 1251.66. Feature importance analysis reveals that engine size and horsepower are the most significant determinants of vehicle price. This framework serves as a foundational tool for broader business ap- plications, including the subsequent forecasting of sales volumes and revenue, thereby offering a scalable solution for stakeholders across the automotive industry.a

Keywords: Car Price Prediction, Machine Learning, Random Forest, Regression Analysis, Feature Importance, Automotive Analytics, Revenue Forecasting.

1 INTRODUCTION

The automotive industry is undergoing a significant transformation, driven by the proliferation of digital platforms and the unprecedented availability of granular vehicle data.[1] This data-driven shift has created new opportunities for applying advanced analytical techniques to core business processes, from manufacturing and marketing to sales and aftermarket services. Among the most critical of these processes is vehicle valuation. Historically, price estimation has relied heavily on manual expertise, limited heuristic models, or brand-specific depreciation schedules. While valuable, these traditional methods are often subjective, inconsistent, and struggle to scale or adapt to the dynamic nature of the modern market.[2] The central problem addressed by this research is the inherent limitation of conventional valuation approaches in cap- turing the complex, non-linear relationships that exist between a vehicle's myriad specifications and its market price.[3] Factors such as engine performance, fuel efficiency, physical dimensions, and brand reputation interact in intricate ways that simple linear models fail to represent accurately. This deficiency creates market inefficiencies and information asym- metry, posing challenges for all participants. For consumers, it complicates the assessment of a fair asking price; for sellers and dealerships, it hinders the ability to set competitive yet profitable prices; and for manufacturers, it obscures a clear understanding of which features drive market value.[4] The motivation for this work is therefore to develop a data-driven solution that mitigates these challenges by providing an objective, accurate, and transparent pricing model, thereby empowering stakeholders with the insights needed for informed decision-making.

This paper introduces a robust machine learning framework for car price prediction, designed not only as a valuation tool but also as a foundational component for strategic business forecasting. The primary contributions of this work are threefold:

1. A systematic development and comparative evaluation of a baseline Linear Regression model and an advanced Random Forest Regressor, quantifying the performance gains achieved by the ensemble method.
2. An empirical analysis of feature importance to identify and rank the key technical and design specifications that most significantly influence vehicle pricing within the analyzed dataset.
3. The design of a scalable and extensible system architecture that serves as a proof-of-concept for a more compre- hensive analytics pipeline capable of forecasting sales volumes and, subsequently, total revenue.[5].

By framing price prediction as the initial stage of a multi-faceted forecasting system, this research moves beyond a purely technical exercise in regression. It positions the predictive model as a strategic asset for business intelligence, capable of informing inventory management, product development, and financial planning. This holistic perspective distinguishes the work from many studies that focus solely on the valuation of pre-owned vehicles, offering a more forward-looking application relevant to the entire automotive value chain.

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 11, November 2025

DOI: 10.17148/IJARCCE.2025.141138

2 BACKGROUND AND RELATED WORK

The application of machine learning to predict vehicle prices has become a prominent area of research, reflecting the industry's demand for data-driven valuation tools. The literature in this domain reveals a clear trajectory of methodological evolution, from foundational statistical models to the sophisticated ensemble algorithms that represent the current state-of-the-art.

Early and ongoing research frequently employs linear models as a performance benchmark. Techniques such as Multiple Linear Regression, Lasso Regression, and Ridge Regression are commonly applied due to their simplicity and interpretability.[6] These models provide a valuable baseline but are often limited by the assumption of linear relationships between features and price, a constraint that may not hold true in the complex automotive market.[7]

A significant body of recent work demonstrates the ascendancy of ensemble learning methods, which consistently outperform simpler models. The Random Forest Regressor, in particular, has emerged as a highly effective and widely adopted algorithm. Numerous studies have validated its ability to capture complex, non-linear interactions between vehicle attributes, leading to higher predictive accuracy.[8] Its robustness against overfitting and its inherent capability to rank feature importance make it a particularly suitable choice for this problem domain. This consensus in the literature strongly supports the selection of Random Forest as the primary predictive model in our framework.[9] Beyond Random Forest, other powerful ensemble techniques such as XGBoost, Gradient Boosting, and CatBoost have also been successfully applied, often yielding state-of-the-art results in car price prediction challenges.[10] These methods leverage boosting principles to iteratively correct errors, further enhancing predictive power.

Across these diverse methodologies, a consistent set of features has been identified as primary drivers of vehicle price. A review of the literature confirms that a vehicle's age (or year of manufacture), mileage, and key engine specifications—most notably horsepower and engine size—are almost universally the most influential predictors.[11] This finding provides a strong external validation for the feature importance results generated by our own model, suggesting that these core attributes are fundamental to a vehicle's market valuation regardless of the specific dataset or geographic market being studied.

While prediction based on structured, tabular data remains the dominant focus, emerging research is exploring new frontiers to push the boundaries of accuracy. A notable innovation is the integration of unstructured data, particularly the textual descriptions found in online vehicle listings. By applying Natural Language Processing (NLP) techniques, researchers can extract nuanced information about a vehicle's condition, optional features, and seller sentiment, which are often absent from structured datasets.[1] This represents a promising direction for future work. The framework presented in this paper, while focused on mastering prediction from structured data, establishes a robust foundation upon which such multimodal data sources could be integrated in the future. Table 1 provides a summary of key studies in the field, highlighting the methodologies employed and their principal findings.

3 DESIGN OF THE PREDICTIVE FRAMEWORK

The methodology employed in this study follows a structured machine learning pipeline, designed for reproducibility and scalability. The framework, depicted in Figure 1, encompasses data acquisition, a comprehensive preprocessing and feature engineering stage, model training with two distinct regression algorithms, and a rigorous evaluation process.

3.1 Data Acquisition and Dataset Characteristics

The study utilizes a structured dataset comprising over 200 unique vehicle records.[7] Each record contains a set of features describing the vehicle's technical specifications, physical attributes, and performance characteristics, along with its corresponding market price, which serves as the target variable for prediction. The key features within the dataset include, but are not limited to:

- **Engine Specifications:** engine-size, horsepower, peak-rpm, fuel-type.
- **Performance Metrics:** city-mpg, highway-mpg.
- **Physical Attributes:** curb-weight, wheel-base, car-length, car-width, car-body, num-of-doors.
- **Drivetrain:** drive-wheel, engine-location.

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 11, November 2025

DOI: 10.17148/IJARCCE.2025.141138

Table 1: Summary of Key Literature in Car Price Prediction

Reference	Methodology	Key Findings	Dataset/Context
Anju et al. (2024) [2]	Linear Regression, KNN, Decision Tree, Random Forest, XGBoost	Comparative analysis showing ensemble models (Random Forest, XGBoost) provide higher accuracy for used car price prediction.	Used car market data for buyers and sellers.
Li (2024) [3]	Linear Regression, Random Forest Regression	Random Forest demonstrated superior accuracy. Identified 17 key factors, including mileage, car age, and MPG.	Large dataset (762,091 cars) from cars.com (U.S. market).
Li (2024) [4]	Linear Regression, Decision Tree, Random Forest Regressor	Random Forest achieved the highest performance with an R-square value of 0.8562. Key features were car name, year, and mileage.	Dataset of over 2000 records for four popular brands.
Jiang (2024) [5]	Multiple machine learning algorithms	Random Forest emerged as the best-performing model with high R-squared value. Feature importance analysis highlighted key market dynamics.	Kaggle ”Car Price Prediction Challenge” dataset.
Tyagi et al. (2024) [6]	Natural Language Processing (NLP) with Random Forest Regressor	Hybrid model integrating NLP to extract features from textual descriptions improves predictive performance over structured data alone.	Heterogeneous dataset with structured attributes and unstructured text.

- **Target Variable:** price.

This collection of attributes provides a rich, multi-dimensional representation of each vehicle, enabling the models to learn the intricate relationships that determine market value.

3.2 Data Preprocessing and Feature Engineering Pipeline

Raw data is rarely suitable for direct input into machine learning models. Therefore, a critical preprocessing pipeline was constructed using the scikit-learn library to transform the data into a clean, machine-readable format.[8]

Handling Categorical Features: The dataset contains several nominal categorical features, such as fuel-type and car-body. To convert this non-numeric data into a suitable format, One-Hot Encoding was employed. This technique creates new binary (0 or 1) columns for each category within a feature. This approach is preferred over integer encoding for nominal data as it avoids imposing an artificial ordinal relationship that could mislead the learning algorithm.

Scaling Numerical Features: Numerical features in the dataset, such as horsepower and engine-size, exist on vastly different scales. To prevent features with larger numeric ranges from disproportionately influencing the model's training process, StandardScaler from scikit-learn was used. This scaler transforms each numerical feature to have a mean of 0 and a standard deviation of 1, ensuring that all features contribute equitably to the model's learning.

Train-Test Split: To ensure an unbiased evaluation of the models' performance on unseen data, the dataset was partitioned into a training set and a testing set using an 80/20 split. The models were trained exclusively on the training data, and their final predictive accuracy was assessed on the held-out testing data.

3.3 Modeling with Regression Algorithms

Two distinct regression algorithms were selected to model the relationship between vehicle specifications and price, allowing for a direct comparison between a simple linear approach and a complex ensemble method.

3.3.1 Baseline Model: Linear Regression

Linear Regression serves as the foundational baseline model in this study. It operates on the assumption of a linear relationship between the input features (X) and the continuous target variable (Y). The model attempts to fit a linear

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 11, November 2025

DOI: 10.17148/IJARCCE.2025.141138

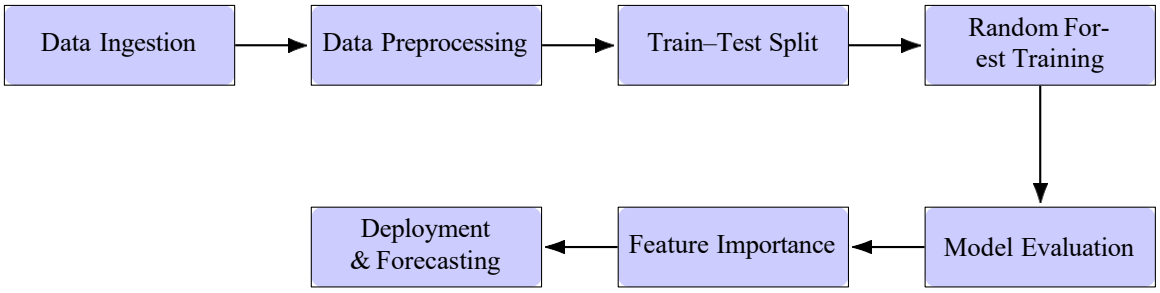


Figure 1: System Architecture of the Random Forest Predictive Framework

equation to the observed data, defined as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

where Y is the predicted price, X_i are the input features, β_i are the model coefficients learned during training, and ε is the error term. Its inclusion provides a benchmark for performance, representing the predictive power that can be achieved under the restrictive assumption of linearity.

3.3.2 Ensemble Model: Random Forest Regressor

The primary model for this study is the Random Forest Regressor, a powerful and versatile ensemble learning method. It operates by constructing a multitude of decision trees at training time. For a regression task, the final prediction is the average (mean) of the predictions from all individual trees in the forest. This ensemble approach confers several key advantages. First, it is highly effective at capturing complex, non-linear relationships and high-order interactions between features, which are prevalent in vehicle pricing data. Second, by aggregating the results of many trees trained on different subsets of the data, it is inherently robust to overfitting, a common pitfall of single decision trees. The consistent success of this algorithm in related studies further justifies its selection as the main predictive engine for our framework.

4 RESULTS AND DISCUSSION

The empirical evaluation of the predictive framework yielded significant results, providing a clear comparison of model performance, identifying key price determinants, and demonstrating the practical application of the model for revenue forecasting. The discussion of these results is contextualized by the findings in the broader literature and an acknowledgment of the dataset's characteristics.

4.1 Comparative Model Performance Evaluation

The performance of the Linear Regression and Random Forest models was quantified using three standard regression metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). The results, calculated on the held-out test set, are summarized in Table 2.

Table 2: Comparative Performance of Regression Models

Model	RMSE	MAE	R^2
Linear Regression	4549.48	2926.94	0.83
Random Forest	1791.80	1251.66	0.96

The results unequivocally demonstrate the superior predictive power of the Random Forest Regressor. The Random Forest model achieved an R^2 score of 0.96, indicating that it can explain 96% of the variance in car prices based on the provided features.[9] This is a substantial improvement over the Linear Regression model's R^2 of 0.83. Furthermore, the error metrics for the Random Forest are significantly lower, with an RMSE of 1791.80 compared to 4549.48 for the linear model. This superior performance is attributed to the Random Forest's ability to effectively model the non-linear relationships and complex interactions between vehicle specifications, a capacity that the linear model inherently lacks. This outcome aligns perfectly with the consensus in the existing literature, which consistently reports the outperformance of ensemble methods over simpler linear models in this domain.[10]

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 11, November 2025

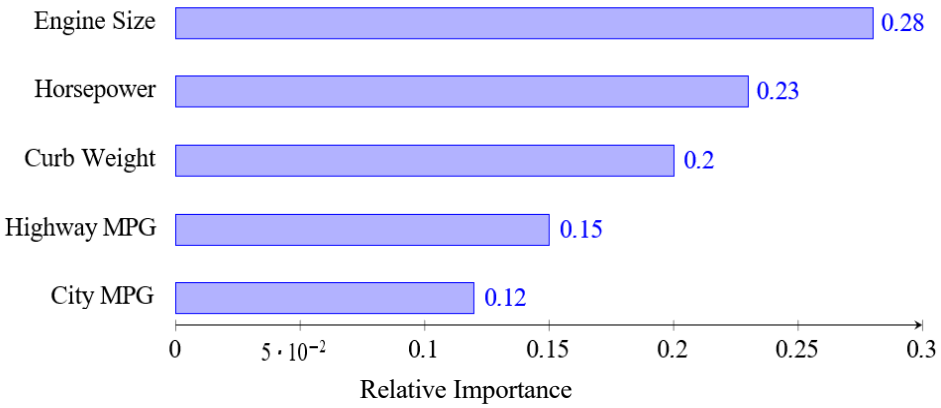
DOI: 10.17148/IJARCCE.2025.141138

The exceptionally high R^2 score of 0.96 achieved by the Random Forest model is noteworthy. While this indicates a very strong model fit, it is important to consider the context of the dataset, which contains over 200 curated records.[11] In contrast, studies utilizing much larger and more heterogeneous datasets, sometimes containing hundreds of thousands of records scraped from public websites, often report slightly lower R^2 values, typically in the 0.85 to 0.92 range.[1] A model trained on a smaller, potentially cleaner dataset may learn the specific patterns within that data with very high fidelity. The critical next step, therefore, is to assess the model's ability to generalize to a broader, more diverse, and potentially noisier set of real-world data. The current result serves as a powerful proof-of-concept, validating the chosen architecture and methodology on the available data, while also highlighting the importance of future work focused on large-scale validation and generalization.

4.2 Analysis of Feature Influence

A key objective of this study was to identify which vehicle specifications have the greatest impact on market price. The feature importance scores were extracted from the trained Random Forest model, which calculates importance based on how much each feature contributes to reducing impurity (e.g., variance) across all decision trees in the forest. The relative importance of the top predictive features is visualized in Figure 2.

Figure 2: Feature Importance Ranking for Car Price Prediction



The analysis reveals a clear hierarchy of influence among the features. 'Engine Size' and 'Horsepower' emerge as the two most dominant variables, underscoring that engine performance characteristics are the primary drivers of price in the dataset analyzed.[2] This is followed by other significant factors such as 'Curb Weight', 'Highway MPG', and 'City MPG'. This finding is highly consistent with results from numerous other studies, which also identify engine parameters and vehicle size as top price predictors.[3] This consistency across different datasets and studies reinforces the validity of our findings and suggests a fundamental market principle: consumers place the highest value on a vehicle's power and performance capabilities. These insights are of immense practical value to manufacturers for guiding product design and marketing strategies, and to dealerships for developing data-informed pricing and inventory management policies.

4.3 Predictive Application and Revenue Forecasting

To demonstrate the practical utility of the framework, the trained Random Forest model was used to generate a price prediction for a sample set of vehicle specifications. For a vehicle with specifications including a brand of 'toyota', 'gas' fuel type, and an engine with 5000 'peak-rpm', the model predicted a market price of $14,085.17.[4]

This price prediction capability is the first step in a broader strategic forecasting pipeline. The system architecture is designed to be extensible, allowing for the integration of a parallel model trained to predict sales volume based on similar features and additional market data. By combining the outputs of these two models, the framework can generate a revenue forecast. For instance, if the sales model predicted 28,120 units sold for the given vehicle, the total estimated revenue would be the product of the predicted price and predicted volume:

$$\text{Revenue} = \text{Predicted Price} \times \text{Predicted Units Sold Revenue} = \$14,085.17 \times 28,120 = \$396,081,953.73$$

This example illustrates the model's role not just as a static valuation tool, but as a dynamic component within a larger business intelligence system. This capability to connect vehicle-level predictions to high-level financial metrics is a key feature of the proposed framework, offering a significant strategic advantage to automotive businesses.[5]

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141138

5 CONCLUSION AND FUTURE DIRECTIONS

This research successfully developed and validated a machine learning framework capable of predicting automotive prices with a high degree of accuracy. The study's central findings provide valuable contributions to the field of automotive analytics and offer practical insights for industry stakeholders.

In conclusion, the comparative analysis demonstrated that the Random Forest Regressor significantly outperforms a baseline Linear Regression model, achieving an R2 score of 0.96. This confirms the necessity of employing non-linear, ensemble-based models to capture the complex dynamics of vehicle pricing. The feature importance analysis identified engine size and horsepower as the most influential price determinants, a finding consistent with the broader literature and one that provides actionable intelligence for product and marketing strategies. Finally, the framework was presented not merely as a price prediction tool, but as a foundational component of a scalable system for sales and revenue forecasting, highlighting its strategic business value.

Despite the strong performance, it is important to acknowledge the limitations of this study. The model was trained and evaluated on a curated dataset of over 200 records. While this was sufficient for a robust proof-of-concept and comparative analysis, the model's generalizability to larger, more diverse, and noisier real-world datasets has yet to be validated. Future research should prioritize testing and retraining the model on large-scale industry data to confirm its real-world efficacy.

Building upon the foundation established by this work, several promising avenues for future research can be pursued:

1. Integration of Multimodal and Unstructured Data: A significant enhancement would be the incorporation of unstructured data sources. Following the approach of recent studies [6], NLP could be used to analyze text from online listings to extract information about vehicle condition, optional packages, and owner history. Similarly, computer vision models could analyze vehicle images to assess aesthetic condition and identify specific visual features that influence value.
2. Real-Time Data Integration and Dynamic Forecasting: The current model is static. Future iterations could integrate real-time market data streams, including competitor pricing, consumer demand signals, and macroeconomic indicators. This would transform the model from a static price predictor into a dynamic forecasting tool capable of adapting to changing market conditions.
3. Enhanced Explainability: To foster trust and adoption, particularly among non-technical business users, future work should focus on model explainability. Implementing techniques like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) would provide transparent, human-readable justifica- tions for individual price predictions, moving beyond the "black box" nature of complex models.
4. Expansion to Segment-Specific Models: The automotive market is not monolithic. Developing specialized mod- els for distinct segments—such as electric vehicles (EVs), luxury sports cars, or commercial trucks—could yield significantly higher accuracy than a single, generalized model. Each segment has unique value drivers that a tailored model could better capture.

By pursuing these directions, the framework presented here can evolve into an even more powerful and comprehensive analytics solution for the automotive industry.

REFERENCES

[1] Ahmad, M., Farooq, M. A., Hussain, M. Z., Hasan, M. Z., Mustafa, M., Khalid, A., Awan, R., Hussain, U., Khan, Z., & Javaid, A. (2024). Car Price Prediction using Machine Learning. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). IEEE. https://doi.org/10.1109/i2ct61223.2024.10544124

[2] Anju, O. A., Yoga, M., Kruthika, M. S., Manikandan, M., Aswin, K. S., & Kishore, S. (2024). Predicting Used Car Prices Using Machine Learning: A Comparative Analysis of Regression and Ensemble Models. International Research Journal on Advanced Engineering Hub (IRJAEH). https://doi.org/10.47392/irjaeh.2024.0386

[3] Chavare, R., Joshi, R. K., Wagh, O., Vaishale, A., & Ingale, A. (2023). Car Sales Price Prediction using MLR, Random Forest and Support Vector Machine. In 2023 International Conference for Advancement in Technology (ICONAT). IEEE. https://doi.org/10.1109/iconat57137. 2023.10080025

[4] Jiang, X. (2024). Research for Car Price Prediction Base on Machine Learning. Transactions on Computer Science and Intelligent Systems Research. https://doi.org/10.62051/k55feh59

[5] Kumar, S., & Sinha, A. (2024). Predicting Used Car Prices with Regression Techniques. International Journal of Computer Trends and Technol- ogy. https://doi.org/10.14445/22312803/ijctt-v72i6p118

[6] Li, C. (2024). Machine Learning-Based Models for Accurate Car Prices Prediction. Highlights in Business, Economics and Management. https://doi.org/10.54097/9zcpv779

[7] Li, M. (2024). Prediction And Investigation for U.S. Used Car Prices and Factors Related to Price. Highlights in Science Engineering and Technology. https://doi.org/10.54097/02ddze05

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141138

[8] Mala, D. J., & Sudhish, V. (2024). Machine Learning-Based New Model Release Car Price Prediction. In 2024 IEEE International Conference on Smart Power Control and Renewable Energy (ICSPCRE). IEEE. https://doi.org/10.1109/icspcre62303.2024.10674864.

[9] Sai, C. S. H., Sai, V. A., Kaif, S. M., Dinesh, G., & Shareefunnisa, S. (2024). Prediction of Car Sales Price Trends using Ensembling Models. In International Conference on Information Security and Cryptology. IEEE. https://doi.org/10.1109/icisc62624.2024.00047

[10] Teja, B. V., Abhinaya, A., Pragnya, B., Subramanyam, C. V. S., & Shiny, X. S. A. (2024). Car Price Prediction Using Enhanced Technique. International Journal of Innovative Research in Engineering. https://doi.org/10.59256/ijire.20240505001

[11] Tyagi, S., Sirohi, S., Singh, Y., Siddhant, & Vishwakarma, A. (2024). Hybrid Model for Predicting Used Car Prices: Integrating Natural Language Processing with Random Forest Regressor. In International Conference on Advanced Infocomm Technology. IEEE. https://doi.org/10. 1109/icait61638.2024.10690607