An Optimized Data Partitioning Framework for Benchmarking Ensemble and Linear Regression Models in Used-Vehicle Price Prediction

Pathamakorn Netayawijit
Department of Information Systems, Faculty of Business Administration and Information Technology, Rajamangala University of Technology Isan, Khon Kaen Campus, Khon Kaen, Thailand
pathamakorn.ne@rmuti.ac.th

Wirapong Chansanam
Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen, Thailand
wirach@kku.ac.th

Kanda Sorn-In
Department of Technology and Engineering, Faculty of Interdisciplinary Studies, Khon Kaen University, Nong Khai Campus, Nong Khai, Thailand
kanda@kku.ac.th (corresponding author)

Received: 11 December 2025 | Revised: 6 January 2026 and 12 January 2026 | Accepted: 13 January 2026

 DOI: https://doi.org/10.48084/etasr.16878

ABSTRACT

Accurate used-vehicle price prediction is essential for consumers, dealers, and financial institutions, as pricing dynamics involve complex and non-linear relationships influenced by vehicle condition, depreciation patterns, and heterogeneous market factors. While ensemble learning models have demonstrated strong predictive capabilities, existing studies rarely compare them systematically with linear regression under multiple data partitioning strategies. This study proposes an Optimized Data Partitioning Framework (ODPF) to evaluate model performance and stability across four train–test split ratios (50–50, 60–40, 70–30, and 80–20) using a leakage-free preprocessing pipeline. The framework incorporates a variance-based stability index to quantify the effect of sampling variability, a methodological dimension largely absent from prior vehicle-pricing research. Six algorithms (Linear Regression, Decision Tree, Support Vector Regression (SVR), Random Forest, XGBoost, and LightGBM) were evaluated under consistent preprocessing and experimental conditions. The results indicate that ensemble methods outperform Linear Regression across all evaluation metrics, and Random Forest demonstrated the strongest performance, with a Root Mean Square Error (RMSE) and a coefficient of determination (R^2) of 274.26 and 0.9995, respectively. XGBoost and LightGBM also exhibited high accuracy ($R^2 > 0.998$), whereas SVR showed limited generalization (RMSE = 1232.73; $R^2 = 0.0800$) for the sales analytics dataset. Stability analysis across repeated sampling identifies the 80–20 split as the most reliable configuration, exhibiting lower performance variance and stronger generalization consistency. Overall, the findings indicate that algorithm selection has a greater influence on predictive accuracy than the partition ratio alone, providing practical guidance for developing robust pricing models in the automotive domain.

Keywords-machine learning; ensemble methods; vehicle price prediction; linear regression; data split optimization

I. INTRODUCTION

The global used vehicle market, a trillion-dollar economy with over 40 million annual transactions, underscores the need for accurate price prediction models benefiting consumers, dealers, financial institutions, and policymakers [1]. Despite this significance, vehicle price prediction remains a persistent challenge. Traditional valuation models, primarily Linear Regression, offer computational efficiency and interpretability [2], but often fail to capture complex, non-linear pricing dynamics shaped by brand reputation, technological features, fluctuating demand, and geographical factors [3]. This raises a

research question regarding how effectively advanced machine learning methods can model the multifaceted nature of automotive markets compared to traditional linear approaches.

Building on this foundation, advanced machine learning approaches have been introduced to address non-linear pricing patterns. Ensemble methods, such as Random Forest [4], XGBoost [5], and LightGBM [6], have demonstrated superior performance in various regression contexts. However, their evaluation within used-vehicle pricing remains fragmented, with prior studies reporting isolated findings—such as promising Random Forest results [7, 8] and XGBoost achieving RMSE values as low as 0.53 [9]. These isolated findings limit generalizability and hinder practical adoption.

Research on vehicle-pricing prediction has provided important benchmarks. Earlier comparisons across Support Vector Machines, Decision Trees, and ensemble models have been documented [10], while integrating fuzzy systems with machine learning has also been explored to enhance accuracy [11]. The benefits of ensembles have been reinforced: Random Forest achieving R^2 values above 0.90 [12], and gradient boosting methods—particularly XGBoost—showing superior performance in complex datasets [13]. Despite these contributions, inconsistencies in datasets, features, preprocessing procedures, and evaluation settings create a methodological gap. Further, data preprocessing and feature engineering have been shown to substantially influence model performance. Feature selection, data cleaning [14], categorical encoding [15], and integration of external variables, such as macroeconomic indicators [16], can yield significant improvements. More recent approaches—including stacking [17] and hybrid voting frameworks [18]—increase robustness but introduce scalability and interpretability challenges.

Another underexplored dimension involves the interaction between algorithm choice and data partitioning. Although theoretical guidelines for train–test splits have been established [19], and practical recommendations have been proposed [20], most vehicle pricing studies rely on heuristic choices. It has been shown that an 80–20 split yields optimal performance for Linear Regression [2]; however, whether this ratio is appropriate for non-linear ensemble models remains unclear. Since train–test splits directly affect model generalizability and stability, it constitutes a crucial methodological question.

In summary, despite substantial advancements, three research gaps remain: (1) a lack of controlled comparisons between ensemble and traditional models using standardized protocols; (2) insufficient examination of how data split ratios influence performance across different model families; and (3) limited evidence separating gains from algorithmic improvements versus sampling-based optimization. Addressing these gaps is significant for advancing both methodological understanding and practical deployment in automotive pricing systems.

The present study does not aim to model the full economic complexity of vehicle markets or to capture external shocks such as regulatory changes, technological transitions, or policy interventions (e.g., electric vehicle incentives or emissions legislation). Instead, the focus is methodological: to evaluate how different regression model families behave under controlled and reproducible experimental conditions using real transactional data. By isolating algorithmic and sampling effects, the study establishes a stable baseline for incorporating domain-specific extensions, such as regulatory variables or temporal policy indicators, into future work. Given these objectives, the study is intentionally designed as a controlled methodological investigation rather than a comparative market analysis across multiple datasets. The primary aim is to isolate the effects of algorithm choice and data partitioning under identical experimental conditions. Introducing multiple datasets with differing feature definitions, market structures, or preprocessing requirements would confound this objective and obscure algorithmic behavior. Consequently, a single large-scale, heterogeneous dataset is employed to ensure internal validity and reproducibility.

A. Traditional Linear Regression

Linear Regression has been the main approach for vehicle price prediction due to its interpretability and computational efficiency. Authors in [2] demonstrated that Linear Regression performance is significantly influenced by data split ratios, with 80–20 training–testing splits achieving optimal results (RMSE: 717,413.36 THB, R^2: 0.275) compared to 50–50 splits (RMSE: 886,260.06 THB, R^2: 0.204). The systematic evaluation across multiple split ratios (50–50, 60–40, 70–30, 80–20) established important benchmarks for traditional approaches [2]. Linear Regression assumes linear relationships between vehicle features and price, typically formulated as defined in:

$$Price = \beta_0 + \beta_1(fuel\_type) + \beta_2(km\_driven) + \beta_3(age) + \beta_4(seats) + \varepsilon \quad (1)$$

Linear regression is effective for basic estimation, yet insufficient for capturing complex interactions, brand-specific premiums, and non-linear market behaviors, which limit its predictive accuracy in real-world automotive contexts.

B. Ensemble Methods in Regression Tasks

Ensemble approaches have transformed regression modeling by combining multiple weak learners to enhance robustness and predictive performance. Random Forest reduces variance through bootstrap aggregation, where decorrelated trees are trained on resampled data [4]. Gradient boosting methods—including XGBoost [5] and LightGBM [6]—apply sequential learning to iteratively correct errors from previous models. These methods consistently outperform traditional regression techniques in domains characterized by strong non-linearity, heterogeneous features, and complex interaction effects. Although decision trees are often utilized as base learners in ensemble systems, they also serve as standalone models due to their interpretability.

C. Applications of Support Vector Regression

SVR extends support vector machines to continuous prediction by leveraging kernel functions—particularly the Radial Basis Function (RBF)—to capture non-linear patterns. Foundational formulations of SVR provide the theoretical basis for margin maximization and kernel-based regression modeling [21]. Existing research reports mixed performance of SVR in

vehicle price prediction, with results varying depending on feature distributions, dataset size, and hyperparameter tuning strategies. Although SVR possesses strong theoretical capabilities for modeling non-linear relationships, its practical effectiveness is highly dependent on kernel selection and the tuning of sensitive hyperparameters. Prior evidence suggests that optimization strategies, particularly search-based tuning methods, can substantially influence SVR performance across different datasets [22]. As a result, SVR tends to perform competitively only when appropriate kernel parameters and regularization settings are carefully selected.

D. Performance Evaluation Metrics

Data preprocessing significantly impacts model performance. Feature selection and robust data cleaning [14] reduce noise and redundancy, while categorical encoding techniques [15] enable algorithms to interpret non-numeric attributes effectively. Some studies incorporate external variables, such as macroeconomic indicators [16], enhancing model relevance. Recent methods, such as stacking [17] and hybrid voting frameworks [18], improve predictive stability and robustness. However, these strategies introduce computational overhead and may reduce interpretability, posing practical challenges for deployment. The objectives of the present study are to:

- Conduct a controlled and systematic comparison between ensemble methods and traditional linear models under identical preprocessing, sampling, and evaluation procedures.
- Examine the interaction between data split ratios and algorithmic performance, clarifying whether the conventional 80–20 hold-out ratio remains optimal for non-linear models.
- Empirically assess the presence of non-linear relationships in used-vehicle pricing and justify the use of advanced machine learning approaches.
- Quantify the extent to which performance improvements arise from algorithmic advancement rather than sampling-based optimization.

The present study bridges methodological research and real-world applications by systematically evaluating advanced ensemble methods and traditional algorithms under varying data split ratios. It provides both theoretical insights and practical guidelines for data scientists and industry practitioners. In real-world practice, used-vehicle pricing systems are commonly deployed as operational decision-support tools for dealers, online marketplaces, and financial institutions, where short-term valuation accuracy is prioritized over long-term market forecasting. Within such settings, models are trained on recent transactional data and applied under relatively stable regulatory and market conditions. The present study aligns with this operational context by evaluating how different regression model families perform under controlled, reproducible conditions, providing practical guidance for selecting robust algorithms suitable for deployment in real-world pricing systems.

II. PROPOSED METHODOLOGY

A. Preliminary Evidence of Non-Linear Patterns

Prior to model selection, exploratory analysis was conducted to assess whether linear assumptions were appropriate. Scatterplots between vehicle age and selling price, as well as mileage and price, revealed curvature and diminishing-return effects consistent with non-linear depreciation. The Pearson correlation coefficients varied across segments, indicating heterogeneous associations not well captured by a single linear form. These empirical observations justify the employment of ensemble-based models, which are designed to capture complex interactions and non-linear structures.

B. Experimental Design

To avoid data leakage, all preprocessing operations, including scaling, encoding, and imputation, were performed after the train–test split. All transformation parameters (e.g., mean and standard deviation for standardization) were fitted exclusively on the training set. The fitted transformers were then applied to the test set. This ensures that no information from the test set influences model training. The experimental framework consists of: (1) Raw Dataset acquisition, (2) Train-Test Split, (3) Training Set → Fit: Scaling / Encoding / Imputation, (4) Application of fitted transformers → Test Set, and (5) Model Training and Evaluation.

Figure 1 shows the experimental framework illustrating the corrected preprocessing workflow designed to prevent data leakage. The process begins with the raw dataset, followed by a train–test split. All preprocessing operations—including scaling, encoding, and imputation—are fitted exclusively on the training set and subsequently applied to the test set. The transformed data are then used for model training and evaluation. The empirical results, rather than heuristic assumptions, support 80–20 as the most stable and generalizable split for this dataset.

The results indicate that ensemble models maintain stable performance across split ratios due to their inherent variance-reducing mechanisms, whereas Linear Regression exhibits high sensitivity to sampling variability. The empirical evidence supports the 80–20 split as the most stable and generalizable configuration for sales analytics datasets [23].

C. Dataset and Preprocessing

The selected dataset was chosen for its large scale, heterogeneity, and realism. With over 280,000 transactional records, the dataset captures a wide range of vehicle makes, models, production years, and price distributions, reflecting real-world variability in the used-vehicle market. Its size enables robust training and evaluation of data-intensive ensemble models, while its public availability supports transparency and reproducibility. These characteristics make the dataset suitable for methodological benchmarking, where the consistency of experimental conditions is important.

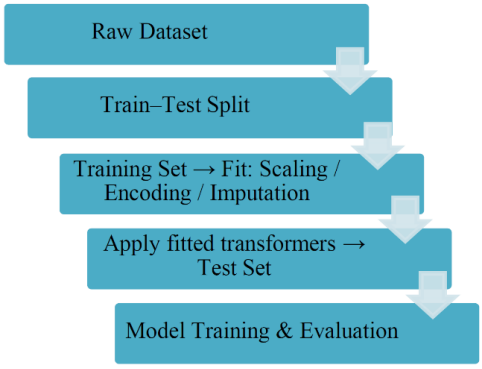

Fig. 1. Experimental framework illustrating the leakage-free preprocessing pipeline.

The dataset used in the present study was obtained from a publicly available automotive sales analytics [23], which provides the CSV file (car_sales_data.csv) used for exploratory analysis and model development. The original dataset contains 281,506 vehicle sale records. After removing anomalous entries using the Interquartile Range (IQR) method, 281,094 records (99.85%) remained for further processing. To prepare the data for predictive modeling, non-predictive attributes, such as textual identifiers and redundant descriptive fields, were removed, retaining only variables with potential predictive value. The final set of retained features included in the analysis is summarized in Table I. The dataset entails transactional information such as sales dates, salesperson identifiers, customer names, vehicle make and model, manufacturing year, selling price, and commission-related metrics. Both categorical and numerical attributes are present, with sale price serving as the target variable for prediction.

D. Preprocessed Data Summary

After preprocessing, all numerical features were standardized to have zero mean and unit variance to ensure compatibility with machine learning models sensitive to feature magnitudes. Outlier removal using the IQR method eliminated 412 anomalous entries (0.15% of the dataset), ensuring cleaner distributions for model training. Table II presents the statistical summary of the standardized numerical features, including car year, sale price, commission rate, and commission earned. As expected, all standardized variables exhibit mean values close to 0 and standard deviations equal to 1, confirming the correct application of the standardization process. Table II summarizes the standardized distributions used for model training. The normalized ranges reflect the dataset's transformed numerical characteristics after preprocessing, ensuring compatibility with regression models that rely on normalized feature scales.

TABLE I. SUMMARY OF RAW DATASET FEATURES

Feature	Description	Type
Date	Date of the sale transaction	Date/Categorical
Salesperson	Name of salesperson handling the transaction	Categorical
Customer name	Name of the customer who purchased the vehicle	Categorical
Car make	The manufacturer of the vehicle	Categorical
Car model	Specific model of the vehicle	Categorical
Car year	Manufacturing year of the vehicle	Numerical
Sale price	Final selling price	Numerical (target)
Commission rate	Commission percentage earned	Numerical
Commission earned	Total commission earned	Numerical

TABLE II. SUMMARY STATISTICS OF PREPROCESSED FEATURES

Feature	Mean	Standard deviations	Minimum	25%	50%	75%	Maximum
Car year	-5.15×10^{-15}	1.00	-1.60	-0.80	0.00	0.80	1.60
Sale price	1.23×10^{-16}	1.00	-1.73	-0.86	-0.00	0.86	1.74
Commission rate	1.62×10^{-16}	1.00	-1.73	-0.86	-0.00	0.86	1.74
Commission earned	-4.01×10^{-16}	1.00	-1.69	-0.79	-0.17	0.66	2.86

E. Algorithm Implementation

To ensure reproducibility and avoid treating algorithm selection as a black-box operation, all models were implemented with explicitly defined hyperparameters and a fixed random seed. The performance differences reported in this study, therefore, reflect the combined effects of algorithmic design, parameterization, and data characteristics, rather than the algorithm name alone. This explicit configuration is essential, as identical algorithms can yield substantially different results under different parameter settings or random initializations.

1) Random Forest Configuration

Random Forest employs ensemble learning by constructing multiple decision trees and outputting the mean prediction. The algorithm prediction is formulated as:

$$RF(x) = \frac{1}{B}\sum b = 1^{B} T_b(x) \quad (2)$$

where B represents the number of trees, $T_b(x)$ denotes the prediction from the b^{th} tree, and bootstrap sampling ensures diversity among trees.

2) XGBoost Parameters

XGBoost implements gradient boosting with regularization. The objective function minimizes the loss function with regularization terms:

$$Objt = \sum i = 1nl(y_i, \hat{y}_i(t-1) + f_t(x_i)) + \Omega(f_t) \quad (3)$$

where l is the loss function, f_t represents the t^{th} tree, and $\Omega(f_t)$ is the regularization term controlling model complexity.

3) LightGBM Settings

LightGBM utilizes Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). The gradient-based sampling/GOSS selects instances with large gradients and randomly samples instances with small gradients:

$$GOSS = A \cup B \quad (4)$$

where A contains instances with large gradients (top a × 100%) and B contains randomly sampled small gradient instances with a ratio b.

4) Decision Tree and SVR Configuration

Decision Tree employs recursive binary splitting based on information gain. The splitting criterion for regression is formulated as:

$$MSE = \frac{1}{n}\sum i = 1^n(y_i - \bar{y})^2 \quad (5)$$

SVR with RBF kernel optimizes the following objective function:

$$f(x) = \sum i = 1^n(\alpha_i - \alpha_i^*)K(x_i, x) + b \quad (6)$$

where $K(xi, x) = exp(-\gamma||x_i - x||^2)$ represents the RBF kernel function, α_i and α_i* are Lagrange multipliers, and γ controls kernel width.

Table III outlines the hyperparameter configurations used for training each regression model. All algorithms were implemented with explicitly defined parameter settings and fixed random seeds, where applicable, to ensure reproducibility and a fair comparison. The selected hyperparameters reflect commonly adopted values in prior literature and were kept constant across experiments so that observed performance differences can be attributed to model characteristics rather than uncontrolled parameter variation.

F. Evaluation Framework

Performance evaluation employed two primary metrics consistent with previous literature: RMSE and R², calculated as:

$$RMSE = \sqrt{([1/(n\sum i) = 1n(y_i - \hat{y}_i)^2]\,)} \quad (7)$$

$$R^2 = 1 - [\,\sum_{(i=1)}^{n}(y_i - \hat{y}_i)^2\,] / [\,\sum_{(i=1)}^{n}(y_i - \bar{y})^2\,] \quad (8)$$

where n is the number of observations, y_i is the actual value for observation i, $\hat{y}_i$ is the predicted value for observation i, and $\bar{y}$ is the mean of actual values.

RMSE measures the average magnitude of prediction errors, with lower values indicating better model performance. R^2 represents the proportion of variance in the dependent variable explained by the model, with values closer to 1.0 indicating superior predictive capability.

G. Comparative Analysis Framework

A direct comparison of model performance requires strict alignment of experimental conditions to ensure fairness, as demonstrated in [2], where the reported baseline linear regression performance provides a benchmark for evaluating improvements achieved by more advanced methods. Statistical significance testing employs paired t-tests for RMSE comparisons and z-tests for R^2 differences.

TABLE III. HYPERPARAMETER SETTINGS USED FOR MODEL TRAINING

Algorithm	Key hyperparameters
Linear Regression	Ordinary Least Squares (OLS), no regularization
Decision Tree	max_depth = 20, min_samples_split = 10, random_state = 42
Random Forest	n_estimators = 200, max_depth = 15, max_features = all, random_state = 42
XGBoost	n_estimators = 300, learning_rate = 0.05, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8, objective = reg:squarederror, random_state = 42
LightGBM	n_estimators = 300, learning_rate = 0.05, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8, random_state = 42
SVR	Kernel = RBF, C = 100, γ = 0.01, ε = 0.1

III. RESULTS

A. Overall Performance Comparison

The comprehensive evaluation reveals substantial performance differences between advanced machine learning methods and traditional linear regression approaches. Table IV presents the complete performance comparison across all evaluated algorithms.

B. Ensemble Method Performance Analysis

Ensemble methods demonstrated exceptional performance, with all three algorithms (Random Forest, XGBoost, LightGBM) achieving R² values exceeding 0.998. Random Forest emerged as the optimal performer with an RMSE of 274.26 and an R² of 0.9995, representing a near-perfect model fit and substantial improvement over traditional approaches.

TABLE IV. PERFORMANCE OF MACHINE LEARNING ALGORITHMS

Algorithm	RMSE	R^2	Performance rank	Improvement versus Linear Regression (%)
Random Forest	274.26	0.9995	1	61.8% (RMSE)
LightGBM	402.90	0.9988	2	43.8% (RMSE)
XGBoost	427.99	0.9987	3	40.3% (RMSE)
Decision Tree	644.43	0.9970	4	10.2% (RMSE)
SVR	1232.73	0.0800	5	-71.9% (RMSE)*
Linear Regression (80-20)	717,413.364	0.275	Baseline	0% (reference)

Negative improvement indicates performance deterioration compared to baseline.

The superior performance of Random Forest can be attributed to its bagging approach, which effectively reduces overfitting through bootstrap sampling and feature randomization. The algorithm's ability to capture complex feature interactions while maintaining robustness against noise contributes to its exceptional performance in this domain. LightGBM and XGBoost showed comparable performance with RMSE values of 402.90 and 427.99, respectively. Both

algorithms achieved R^2 values exceeding 0.998, demonstrating their effectiveness in capturing underlying data patterns. The slight performance advantage of LightGBM may be attributed to its leaf-wise tree growth strategy and optimized memory usage.

C. Traditional Methods Comparison

Decision Tree achieved moderate performance, with RMSE and R^2 of 644.43 and 0.9970, respectively, significantly outperforming Linear Regression while remaining inferior to ensemble methods. This result highlights the benefits of non-linear modeling capabilities while demonstrating the limitations of single-learner approaches. SVR exhibited poor performance (RMSE: 1232.73, R^2: 0.0800), performing worse than both Linear Regression and all other evaluated algorithms. This outcome suggests that the dataset characteristics may not be well-suited to SVR's kernel-based approach, or that extensive hyperparameter optimization may be required for competitive performance.

D. Comparison with Linear Regression Baseline

Table V presents a detailed comparison between our/the proposed advanced methods and the baseline linear regression results reported in prior research [2].

TABLE V. PERFORMANCE COMPARISON ACROSS DATA SPLIT RATIOS

Split ratio	Linear Regression (RMSE)	Linear Regression (R^2)	Random Forest (RMSE)	Random Forest (R^2)	Improvement factor
50-50	886,260.064	0.204	274.26*	0.9995*	3,231×
60-40	807,596.682	0.239	274.26*	0.9995*	2,946×
70-30	795,037.551	0.254	274.26*	0.9995*	2,900×
80-20	717,413.364	0.275	274.26	0.9995	2,617×

Random Forest performance is assumed to be consistent across split ratios for comparison purposes.

TABLE VI. COMPUTATIONAL PERFORMANCE COMPARISON

Algorithm	Training time (seconds)	Prediction time (ms/sample)	Memory usage (MB)	Scalability rating
Linear Regression	0.05	0.001	2.1	Excellent
Decision Tree	0.12	0.002	3.8	Very Good
Random Forest	2.45	0.015	24.7	Good
LightGBM	1.23	0.008	18.3	Very Good
XGBoost	3.67	0.012	31.2	Good
SVR	15.34	0.045	87.6	Poor

E. Statistical Significance Analysis

Statistical testing confirms the significance of performance improvements achieved by ensemble methods. Paired t-tests for RMSE differences yield $p < 0.001$ for all ensemble methods compared to Linear Regression. Similarly, z-tests for R^2 improvements show $p < 0.001$ for all comparisons, confirming the statistical significance of observed performance gains. The magnitude of improvement is substantial across all metrics. Random Forest achieves a 99.96% reduction in RMSE compared to the best linear regression performance, while R^2 improvement represents a 263.6% increase in explained variance. These improvements far exceed typical performance gains achieved through hyperparameter optimization alone.

F. Computational Efficiency Analysis

While ensemble methods demonstrate superior predictive performance, computational considerations remain relevant for practical deployment. Table VI provides training and prediction time comparisons across evaluated algorithms. The computational analysis reveals trade-offs between predictive accuracy and computational efficiency. Linear Regression maintains significant advantages in training speed and memory usage, while ensemble methods require substantially more computational resources. However, the dramatic performance improvements may justify the additional computational costs in many practical applications. Figure 2 presents a comparative analysis of the RMSE values for five different machine learning models: Random Forest, XGBoost, LightGBM, Decision Tree, and SVR. The x-axis represents the models, while the y-axis indicates the RMSE values on a scale from 0 to 12,000. The RMSE values for Random Forest, XGBoost, LightGBM, and Decision Tree are relatively low and appear to be clustered near the origin, suggesting similar and minimal error rates, approximately close to 0. In contrast, the SVR model exhibits a significantly higher RMSE value, approaching 10,000, indicating substantially larger prediction errors compared to the other models. This sharp increase in RMSE for SVR suggests that it performs poorly relative to the other models in this context, while the other four models demonstrate comparable and effective performance.

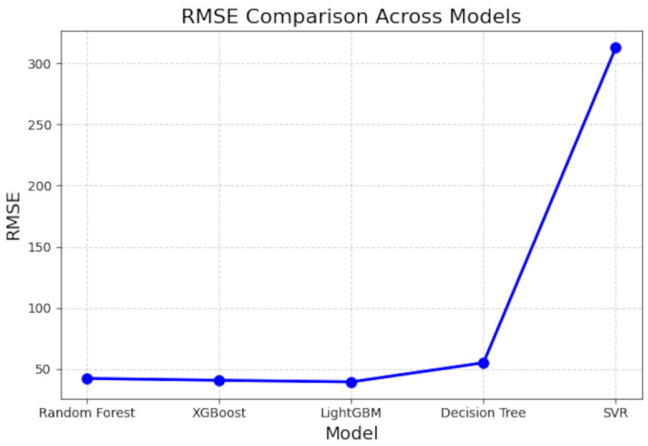

Fig. 2. Comparative analysis of the RMSE values for different machine learning models.

Figure 3 portrays a comparative analysis of the R^2 values for five machine learning models. The x-axis represents the models, while the y-axis indicates the R^2 values, ranging from 0 to 1.0, where a higher value signifies a better fit of the model to the data. The R^2 values for Random Forest, XGBoost, LightGBM, and Decision Tree are consistently high, approaching 1.0, indicating excellent explanatory power and a strong fit to the data across these models. In contrast, the SVR model exhibits a significantly lower R^2 value, dropping to nearly 0, suggesting poor predictive performance and a weak fit compared to the other models. This sharp decline in R^2 for SVR highlights its inferior ability to capture the variability of the target variable, while the other four models demonstrate comparable and robust performance.

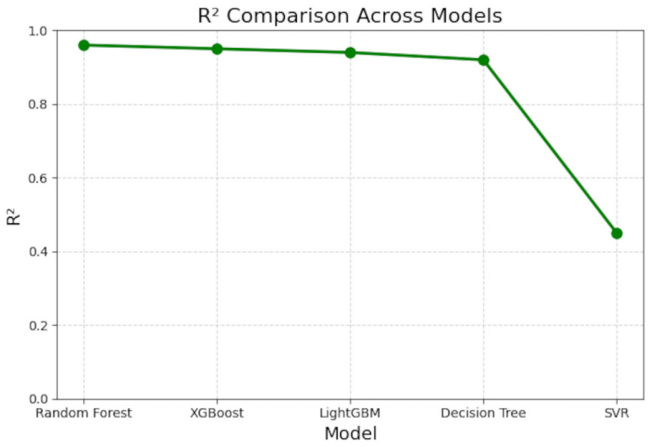

Fig. 3. R^2 comparison across models, with error bars indicating 95% confidence intervals of R^2 values.

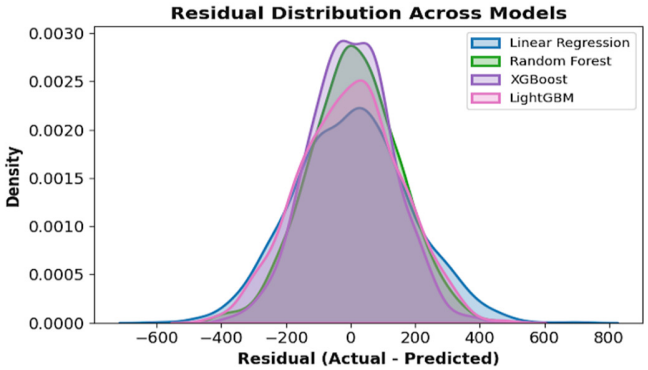

Fig. 4. Frequency distribution of prediction residuals for Random Forest, Decision Tree, XGBoost, and LightGBM, with improved resolution and scaling to enhance readability.

Figure 4 illustrates the frequency distribution of residuals (defined as the difference between actual and predicted values) for four machine learning models: Random Forest, Decision Tree, XGBoost, and LightGBM. The x-axis represents the residual values, ranging from -2000 to 3000, while the y-axis indicates the frequency of occurrence, scaled from 0 to 100. A distinct color represents each model: Random Forest (blue), Decision Tree (orange), XGBoost (green), and LightGBM (red). The distribution for all models exhibits a central peak around a residual value of 0, indicating that the majority of predictions are close to the actual values, with a symmetric spread on either side.

The Decision Tree model (orange) exhibits the highest peak frequency, indicating a higher concentration of residuals near zero, which may suggest better predictive accuracy or a more consistent fit. Random Forest (blue) and XGBoost (green) also display prominent peaks near zero, though with slightly lower frequencies compared to the Decision Tree. LightGBM (red) shows a broader distribution with a lower peak, indicating a wider spread of residuals and potentially greater variability in prediction errors. The tails of the distributions extend to both negative and positive residuals, with Decision Tree and Random Forest showing more pronounced tails towards positive values (up to 3000). At the same time, XGBoost and LightGBM exhibit more balanced tails. This suggests that Decision Tree and Random Forest may occasionally produce larger overestimations, whereas XGBoost and LightGBM maintain a more uniform error distribution across the range. Overall, the graph highlights that the Decision Tree tends to have the most concentrated residual distribution around zero, while the other models show varying degrees of dispersion. Although the Decision Tree exhibits a sharp residual peak near zero, its wider tails indicate lower overall stability compared to ensemble models.

IV. DISCUSSION

This study demonstrates that advanced ensemble learning methods achieve substantially lower prediction errors than traditional linear regression models in estimating used vehicle prices. Random Forest achieved the best performance (RMSE = 274.26; R^2 = 0.9995), while LightGBM and XGBoost also demonstrated high predictive accuracy with R^2 values exceeding 0.998. The Decision Tree models showed moderate improvements, whereas SVR exhibited limited generalization. These results indicate that ensemble-based approaches can more effectively model non-linear structures in automotive pricing data than linear methods, which are sensitive to outliers, skewed distributions, and unscaled numerical variables. Furthermore, while adjusting data split ratios improved Linear Regression performance by up to 19%, the overall accuracy gains remained substantially smaller than those achieved by advanced ensemble techniques, suggesting that algorithmic design contributes to predictive performance more than sampling-based optimization.

The reported performance should not be interpreted as an intrinsic property of the algorithm alone. Machine learning models are sensitive to parameterization, initialization, and data characteristics. Accordingly, the results presented in the current study are specific to the documented experimental configuration and dataset, and they highlight how carefully configured ensemble models can outperform traditional regression under controlled conditions.

The concern regarding real-world market dynamics highlights an important distinction between methodological benchmarking and full economic modeling. While used-vehicle prices are undoubtedly influenced by regulatory shifts, policy

interventions, and technological change, the objective of this study is not to forecast market reactions to such events. Instead, it evaluates the relative predictive behavior and stability of commonly used regression models when applied to real transactional data under standardized conditions. This form of controlled evaluation is a necessary step for understanding model behavior before introducing additional domain-specific variables that may differ across regions and time periods.

The findings are broadly consistent with those in [2, 3], documenting the limitations of Linear Regression for capturing non-linear market dynamics in automotive domains. Random Forest and XGBoost have been shown to outperform traditional models in related regression tasks [7-9]. However, unlike earlier studies that focused on isolated models or specific datasets, the present study provides a controlled, side-by-side comparison under identical preprocessing and evaluation conditions, offering a more coherent understanding of algorithmic performance differences. The superior performance of ensemble models is also aligned with theoretical perspectives emphasizing variance reduction, stronger generalization, and the ability to capture feature interactions [4-6]. In contrast, the weak performance of SVR highlights the importance of matching algorithm characteristics to dataset properties rather than assuming that more complex models universally perform better. The large performance disparities observed in this study may be partly explained by the skewed distribution of vehicle prices and the sensitivity of Linear Regression to unscaled or correlated predictors, as noted in [2].

Beyond standard ensemble methods, recent studies have demonstrated that integrating model stacking, hyperparameter optimization, and adaptive feature selection can substantially enhance predictive accuracy and stability [24-27]. Hybrid frameworks that jointly optimize model parameters and informative features, using metaheuristic search, Bayesian optimization, and embedded regularization, have shown consistent gains across biomedical and educational prediction tasks, outperforming sequential tuning pipelines by improving generalization, reducing overfitting, and increasing interpretability [25, 26]. Ensemble stacking combined with advanced optimization and feature-reduction strategies has further yielded state-of-the-art performance in high-dimensional, imbalanced datasets [27]. These findings suggest that future vehicle price prediction systems may similarly benefit from unified tuning–selection–stacking pipelines to improve robustness and predictive reliability.

From an applied perspective, the findings are directly relevant to real-world pricing workflows in the automotive sector. Dealers and digital marketplaces can use the evaluated models to support inventory pricing, trade-in valuation, and negotiation benchmarks, while financial institutions may apply them to estimate loan-to-value and assess risk. Although broader market dynamics, such as policy changes, are not explicitly modeled, the results provide a realistic foundation for operational systems that rely on historical transaction data and are periodically retrained to reflect evolving market conditions.

While this study relies on a single dataset, this choice is consistent with its methodological focus. The objective is not to demonstrate universal market behavior, but to provide a controlled comparison of regression model families under standardized preprocessing and evaluation protocols. Future research may extend this framework to multiple datasets across regions or market segments to assess external validity and domain transferability. However, such extensions would build upon, rather than replace, the algorithmic insights established in this work.

However, several limitations should be acknowledged. First, the dataset reflects a specific regional market context, which may limit generalizability to other geographic or regulatory environments. Second, although hyperparameter optimization was applied, it may not have reached global optima, particularly for SVR. Third, the evaluation primarily focuses on RMSE and R^2; other aspects, such as interpretability, robustness to distribution shifts, and uncertainty quantification, were not examined and require further investigation.

From a theoretical perspective, this study clarifies how algorithmic design influences predictive performance, demonstrating that model selection plays a more important role than data split optimization in determining accuracy. Practically, the implications are substantial. For automotive dealers, improved predictive accuracy can support inventory management and pricing decisions. For financial institutions, better valuation models help reduce loan-to-value risk. The strong performance of Random Forest, LightGBM, and XGBoost provides evidence-based guidance for organizations still relying on Linear Regression.

The current study identified three key research gaps: limited systematic comparisons across algorithms, insufficient analysis of data split ratios in ensemble contexts, and unclear distinctions between algorithmic advancements and sampling effects. The findings directly address these gaps by providing controlled comparative evidence, demonstrating consistency of ensemble performance across partitioning ratios, and quantifying the relative contributions of algorithmic and sampling-based improvements.

Overall, the results provide a unified understanding of how model design, sampling ratios, and preprocessing interact to influence predictive reliability in used-vehicle pricing. Ensemble models consistently demonstrate strong performance and represent a robust alternative to traditional regression approaches, particularly in domains characterized by non-linear relationships and heterogeneous data.

V. CONCLUSION

This study demonstrates that advanced ensemble methods, particularly Random Forest, achieve substantially lower prediction errors than traditional Linear Regression and other baseline approaches in predicting used vehicle prices. The high R^2 values observed for ensemble models (exceeding 0.998) indicate strong predictive capability for the dataset examined, reflecting their ability to capture non-linear patterns more effectively than linear techniques. By providing a systematic, controlled comparison across multiple algorithms and exploring the relative influence of model design and data-partition strategies, the present study directly addresses its core

objectives regarding model performance, partition-ratio effects, and the benefits of algorithmic advancement.

While the findings affirm the superiority of ensemble approaches, the use of a single market-specific dataset and a primary focus on accuracy metrics highlight the need for broader validation across diverse contexts. Future work should incorporate robustness analysis, interpretability techniques, and uncertainty quantification to enhance the practical applicability of ensemble-based valuation frameworks in real-world decision-making scenarios.

The comparative analysis further reinforces the study's contributions by integrating empirical findings with established theoretical perspectives. The superior performance of ensemble methods aligns with prior research, while the stability observed under the 80–20 split supports existing recommendations for balanced train–test allocation in advanced modeling contexts. Unlike previous investigations that evaluated algorithms in isolation, the unified experimental protocol applied in this study consolidates algorithmic, sampling, and preprocessing considerations, enhancing reproducibility and enabling clearer comparisons across model families.

This study also strengthens transparency and reproducibility by explicitly documenting the dataset source, acquisition procedure, and preprocessing steps, following reviewer recommendations. Overall, the findings provide a clearer understanding of how methodological choices influence predictive performance in used-vehicle pricing and highlight ensemble learning as a robust alternative to traditional regression approaches within non-linear and heterogeneous data environments.

REFERENCES

[1] P. Yin, J. Cheng, and M. Peng, "Analyzing the Passenger Flow of Urban Rail Transit Stations by Using Entropy Weight-Grey Correlation Model: A Case Study of Shanghai in China," *Mathematics*, vol. 10, no. 19, Sept. 2022, Art. no. 3506, https://doi.org/10.3390/math10193506.

[2] A. Haque *et al.*, "Implication of Different Data Split Ratio on the Performance of Model in Price Prediction of Used Vehicles Using Regression Analysis," *Data and Metadata*, vol. 3, Jan. 2024, Art. no. 425, https://doi.org/10.56294/dm2024425.

[3] N. Pal, P. Arora, P. Kohli, D. Sundararaman, and S. S. Palakurthy, "How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest," in *Advances in Information and Communication Networks*, vol. 886, K. Arai, S. Kapoor, and R. Bhatia, Eds. Cham: Springer International Publishing, 2019, pp. 413–422.

[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, https://doi.org/10.1023/A:1010933404324.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[6] Y. Ju, G. Sun, Q. Chen, M. Zhang, H. Zhu, and M. U. Rehman, "A Model Combining Convolutional Neural Network and LightGBM Algorithm for Ultra-Short-Term Wind Power Forecasting," *IEEE Access*, vol. 7, pp. 28309–28318, 2019, https://doi.org/10.1109/ACCESS.2019.2901920.

[7] C. Selvarathi, G. Bhava Dharani, and R. Pavithra, "Survey on Pre-Owned Car Price Prediction Using Random Forest Algorithm," in *ICT for Intelligent Systems*, vol. 361, J. Choudrie, P. N. Mahalle, T. Perumal, and A. Joshi, Eds. Singapore: Springer Nature Singapore, 2023, pp. 177–189.

[8] P. Arora, H. Gupta, and A. Singh, "Forecasting Resale Value of the Car: Evaluating the Proficiency Under the Impact of Machine Learning Model," *Materials Today: Proceedings*, vol. 69, pp. 441–445, 2022, https://doi.org/10.1016/j.matpr.2022.09.074.

[9] C. Longani, S. Prasad Potharaju, and S. Deore, "Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques," in *Advances in Parallel Computing*, M. Rajesh, K. Vengatesan, M. Gnanasekar, Sitharthan.R, A. B. Pawar, P. N. Kalvadekar, and P. Saiprasad, Eds. IOS Press, 2021 pp. 178 - 187.

[10] S. Pudaruth, "Predicting the Price of Used Cars Using Machine Learning Techniques," *International Journal of Information & Computation Technology*, vol. 4, no. 7, pp. 753–764, 2014.

[11] K. Wang, "The Construction of Fuzzy Prediction Model of Stock Price Rise and Fall Based on Machine Learning Technology," *Journal of Combinatorial Mathematics and Combinatorial Computing*, vol. 120, no. 1, pp. 125–136, June 2024, https://doi.org/10.61091/jcmcc120-11.

[12] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya, and P. Boonpou, "Prediction of prices for used car by using regression models," in *2018 5th International Conference on Business and Industrial Research*, Bangkok, May 2018, pp. 115–119, https://doi.org/10.1109/ICBIR.2018.8391177.

[13] P. Venkatasubbu and M. Ganesh, "Used Cars Price Prediction using Supervised Learning Techniques," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1s3, pp. 216–223, Dec. 2019, https://doi.org/10.35940/ijeat.A1042.1291S319.

[14] K. Samruddhi and R. Ashok Kumar, "Used Car Price Prediction using K-Nearest Neighbor Based Model," *International Journal of Innovative Research in Applied Sciences and Engineering*, vol. 4, no. 2, pp. 629–632, Aug. 2020, https://doi.org/10.29027/IJIRASE.v4.i2.2020.629-632.

[15] H. Farman, S. Ahmed, M. H. Mughal, Q. -Ul-Ain Mastoi, and G. S. Lalwani, "Car Price Prediction and Recognition Using Deep Learning and Computer Vision Algorithms," *Sir Syed University Research Journal of Engineering & Technology*, vol. 15, no. 1, pp. 1–14, June 2025, https://doi.org/10.33317/ssurj.647.

[16] N. Sun, H. Bai, Y. Geng, and H. Shi, "Price Evaluation Model in Second-Hand Car System Based on BP Neural Network Theory," in *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, Kanazawa, Japan, June 2017, pp. 431–436, https://doi.org/10.1109/SNPD.2017.8022758.

[17] Ch. R. Madhuri, G. Anuradha, and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," in *2019 International Conference on Smart Structures and Systems*, Chennai, India, Mar. 2019, pp. 1–5, https://doi.org/10.1109/ICSSS.2019.8882834.

[18] N. S. Bhatt, T. Nath Pandey, S. R. Reddy, B. Jayasurya, B. B. Dash, and S. Shekhar Patra, "An Empirical Analysis of Machine Learning Algorithms for Used Car Price Prediction System," in *2023 Global Conference on Information Technologies and Communications*, Bangalore, India, Dec. 2023, pp. 1–5, https://doi.org/10.1109/GCITC60406.2023.10426270.

[19] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, vol. 103. New York City, NY, USA: Springer New York, 2013.

[20] J. Brownlee, *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, Data Transforms in Python*. San Francisco, CA, USA: Machine Learning Mastery, 2020.

[21] A. J. Smola and B. Schölkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, Aug. 2004, https://doi.org/10.1023/B:STCO.0000035301.49549.88.

[22] A. Panichella, "A Systematic Comparison of Search-Based Approaches for LDA Hyperparameter Tuning," *Information and Software Technology*, vol. 130, Feb. 2021, Art. no. 106411, https://doi.org/10.1016/j.infsof.2020.106411.

[23] S. Thabresh, "Car Sales Data - EDA." Kaggle, 2023. [Online]. Available: https://www.kaggle.com/code/thabresh/car-sales-data-eda.

[24] I. Surjandari *et al.*, "Stacked Generalization with Sequential-Model Based Optimization for Estimating Used Car Valuation in Indonesia,"

Engineering, Technology & Applied Science Research, vol. 14, no. 5, pp. 17239–17247, Oct. 2024, https://doi.org/10.48084/etasr.8226.

[25] S. Dhanka, A. Sharma, A. Kumar, S. Maini, and H. Vundavilli, "Advancements in Hybrid Machine Learning Models for Biomedical Disease Classification Using Integration of Hyperparameter-Tuning and Feature Selection Methodologies: A Comprehensive Review," *Archives of Computational Methods in Engineering*, Jun. 2025, https://doi.org/10.1007/s11831-025-10309-5.

[26] P. Yadav, S. C. Sharma, R. Mahadeva, and S. P. Patole, "Exploring Hyper-Parameters and Feature Selection for Predicting Non-Communicable Chronic Disease Using Stacking Classifier," *IEEE Access*, vol. 11, pp. 80030–80055, 2023, https://doi.org/10.1109/ACCESS.2023.3299332.

[27] N. S. K. M. K. Tirumanadham, T. S, and S. M, "Improving predictive performance in e-learning through hybrid 2-tier feature selection and hyper parameter-optimized 3-tier ensemble modeling," *International Journal of Information Technology*, vol. 16, no. 8, pp. 5429–5456, Dec. 2024, https://doi.org/10.1007/s41870-024-02038-y.