An Empirical Evaluation of Regression and Tree-Based Models for Used Car Price Estimation

1st Moustafa Mortada Mohamed
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2023170614@cis.asu.edu.eg

2nd Mohammed Ahmed Mohammed Abdelghani
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2023170477@cis.asu.edu.eg

3rd Ammar Mohammed Ali Ali
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2023170376@cis.asu.edu.eg

4th Omar Karam Sayed Ramadan
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2023170393@cis.asu.edu.eg

5th Salma Khaled Mahmoud Al-Bahai
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2023170258@cis.asu.edu.eg

6th Peter Emad Adly Shafiq
Faculty of Computer and Information Sciences
Ain Shams University
Cairo,Egypt
2023170145@cis.asu.edu.eg

Abstract*—Accurate price estimation in the used car market is a challenging problem due to the complex and nonlinear interactions among vehicle characteristics, particularly in large and diverse markets such as India. This study aimed to investigate and evaluate the effectiveness of statistical and machine learning models for predicting used car prices in the Indian second-hand car market. Specifically, the objectives were to analyze the impact of vehicle attributes on price formation and to compare the predictive performance of Linear Regression, Decision Tree Regression, XGBoost, and Random Forest models under a unified preprocessing and evaluation framework. A real-world dataset containing 9,582 used car listings with 11 attributes, collected up to November 2024, was cleaned and preprocessed using rigorous data preparation techniques, including numerical formatting, categorical encoding, logarithmic transformation of the target variable, and percentile-based outlier treatment. The dataset was then split into training and testing subsets using an 80–20 ratio to ensure unbiased model evaluation. Each model was trained on the same feature set and evaluated using standard regression metrics, including R^2, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The results showed that ensemble-based methods significantly outperformed the baseline linear and single-tree models. In particular, XGBoost and Random Forest achieved the highest predictive accuracy, with R^2 values of approximately 0.85, demonstrating their ability to capture complex nonlinear relationships and feature interactions in used car pricing. These findings highlight the effectiveness of ensemble learning techniques for real-world price prediction tasks and provide a robust foundation for data-driven decision support in the Indian used car market.

I. Introduction

The rapid growth of population and urbanization in India has led to a significant increase in transportation demand. Given the country's large geographical area and diverse economic conditions, personal vehicles play a critical role in daily mobility. However, the high cost of new vehicles has made the used car market an essential alternative for a large segment of the population. As a result, the Indian second-hand car market has expanded rapidly, creating a strong demand for reliable and transparent vehicle price estimation systems. Accurate used car price prediction is a challenging task due to the complex interaction of multiple factors such as vehicle age, mileage, brand, ownership history, fuel type, and transmission. Previous research has shown that traditional statistical models, including linear regression, can capture general pricing trends but often fail to model nonlinear relationships and feature interactions inherent in real-world car pricing data. More advanced machine learning models, such as decision trees and ensemble methods, have demonstrated improved accuracy; however, many existing studies focus on a single model, use limited datasets, or lack a systematic comparison across multiple algorithms under consistent preprocessing and evaluation settings. While existing literature confirms that ensemble learning techniques can outperform simpler models, there remains a gap in comprehensive evaluations using large, real-world datasets from the Indian used car market. Specifically, there is limited empirical analysis comparing linear models, tree-based models, and boosting-based ensembles using the same feature set, preprocessing pipeline, and evaluation metrics. Moreover, few studies investigate model stability, residual behavior, and the practical implications of prediction accuracy in a real market context. To address these gaps, this study investigates

used car price prediction using a dataset collected from the Indian market, comprising 9,582 entries with 11 attributes and covering data up to November 2024. After data cleaning and preprocessing, 9,238 valid records were retained for modeling. The dataset provides a comprehensive view of the second-hand car market in India and enables a robust comparison of predictive models under realistic conditions. The main objective of this paper is to develop and evaluate multiple price prediction models, including Linear Regression, Decision Tree Regression, XGBoost, and Random Forest. The study follows a structured methodology involving data preprocessing, feature transformation, model training, hyperparameter tuning, and performance evaluation using standard regression metrics. By comparing traditional and ensemble-based approaches, this work aims to identify the most effective model for predicting used car prices and to provide insights that can support practical decision-making in automotive market analysis.

II. Model Architecture

The general architecture of the proposed framework consists of four main stages. The first stage focuses on data preprocessing, where the raw used-car dataset is cleaned, transformed, and prepared through numerical formatting, categorical encoding, logarithmic transformation of the target variable, and outlier handling to ensure statistical robustness. In the second stage, model development, multiple predictive models are trained using the processed data, including a baseline Linear Regression model, a Decision Tree regression model, and two ensemble-based methods, XGBoost and Random Forest, to capture both linear and nonlinear relationships in vehicle pricing. The third stage involves model training and evaluation, where the dataset is split into training and testing subsets, and model performance is assessed using standard regression metrics such as R^2, RMSE, and MAE to ensure fair and consistent comparison across models. Finally, the performance analysis stage summarizes the predictive effectiveness of each approach, highlighting the strengths of ensemble models in modeling complex feature interactions and improving generalization performance. This structured architecture enables a systematic comparison of traditional and advanced machine learning techniques for used car price prediction. "'latex

III. Methodology

The dataset was collected from online used-car listings and preprocessed using the `tidyverse` ecosystem in R to ensure suitability for statistical inference and machine learning models. The preprocessing pipeline consisted of the following steps.

A. Data Cleaning and Formatting

- Non-numeric characters such as currency symbols (), mileage units (`km`), and commas were removed from `AskPrice` and `kmDriven`. These variables were then converted to numeric data types.
- Vehicle age was recalculated using $\text{Age} = 2025 - \text{Year}$ to ensure temporal consistency. The original `Year` attribute was removed to avoid multicollinearity.

B. Handling High Cardinality Categorical Features

- **Brand**: The top 10 most frequent brands, representing approximately 85% of the dataset, were retained. All remaining brands were grouped into a single category labeled `Other_Brands`.
- **Model**: The top 50 most frequent models, covering approximately 64% of the dataset, were retained, while the remaining models were grouped into `Other_Models`.

These steps ensured sufficient sample sizes for categorical levels and reduced the risk of overfitting.

C. Statistical Encoding and Transformations

- **Factor Encoding**: Categorical variables (`Brand`, `Model`, `Transmission`, `FuelType`, and `Owner`) were encoded as factors to enable proper handling by regression and tree-based models.
- **Log Transformation**: The target variable `AskPrice` exhibited right-skewness; therefore, a logarithmic transformation was applied, yielding `log_price`. This transformation reduced heteroscedasticity and improved adherence to linear regression assumptions.

D. Outlier Treatment

- **Price**: Observations below the 1st percentile (likely non-functional vehicles) and above the 99th percentile (super-luxury vehicles exceeding 8,000,000) were removed to mitigate the influence of extreme leverage points.
- **Mileage**: Vehicles within the top 1% of mileage values (greater than 220,000 km) were excluded to reduce non-linear depreciation effects associated with end-of-life vehicles.

E. Predictive Models

All predictive models were evaluated using an 80–20 train–test split with a fixed random seed to ensure reproducibility. Categorical variables were encoded consistently as factors or dummy variables based on the requirements of each model.

1) Linear Regression: **Model Choice:** Linear regression was employed as a baseline statistical model to predict `log_price` using both numerical and categorical vehicle attributes.

Data Preparation:

- Numerical features: `km_driven`, `age`
- Categorical features: `brand`, `model`, `transmission`, `fuel_type`, `owner`
- Categorical variables encoded as factors

Model Training:

- Multiple linear regression fitted using the `lm()` function in R
- Model coefficients analyzed to assess feature contributions

Prediction and Back-Transformation: Predictions were generated in logarithmic scale, exponentiated to obtain price estimates, and a smearing factor was applied to correct for retransformation bias.

2) Decision Tree Regression: **Model Choice:** Decision tree regression was selected to capture non-linear relationships and feature interactions.

Model Training:

- Recursive partitioning using the ANOVA splitting criterion
- Splits selected to minimize prediction error at each node

Prediction and Back-Transformation: Predictions obtained in log scale were retransformed to the original price scale.

3) XGBoost Regression: **Model Choice:** Extreme Gradient Boosting (XGBoost) was employed due to its ability to model complex non-linear interactions and assign feature importance.

Data Preparation and Training:

- Categorical variables converted to dummy variables
- Data split using stratified sampling via `createDataPartition`
- Training performed using `xgb.DMatrix` for computational efficiency
- Hyperparameters tuned, including learning rate (`eta`), maximum tree depth, subsampling rate, and column sampling rate
- Early stopping applied based on validation performance

Prediction and Back-Transformation: Predicted log prices were retransformed to the original price scale.

4) Random Forest Regression: **Model Choice:** Random forest regression was utilized for its robustness to multicollinearity and capacity to model complex non-linear patterns.

Model Training:

- Ensemble of decision trees trained on bootstrap samples
- Random feature selection applied at each split to reduce variance

Prediction and Back-Transformation: Predictions generated in logarithmic scale were exponentiated to obtain final price estimates.

IV. Results

A. Preprocessing Results

The preprocessing pipeline reduced data inconsistencies and prepared the dataset for modeling. The impact on the dataset is summarized below:

- Initial Rows: 9,582
- Final Rows: 9,238
- Data Lost: 3.59%

This reduction reflects removal of extreme outliers and invalid entries, ensuring a clean dataset for modeling.

B. Linear Regression

A linear regression model was trained to predict log-transformed car prices using the following predictors: km_driven, age, brand, model, transmission, fuel_type, and owner.

- **R-squared:** 0.7975 (training), 0.7755 (testing)
- **Key predictors:** Age, km_driven, ownership status, and premium brands were statistically significant with expected positive or negative effects.
- **Prediction analysis:** Actual vs. predicted plots indicated that the model effectively captured price trends.

Performance Metrics: The model's evaluation is represented as:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2}$$

$$R^2 = 1 - \frac{SSE}{SST} \tag{3}$$

where y_i is the actual price, $\hat{y}_i$ is the predicted price, and $\bar{y}$ is the mean of actual prices.

Graphical Analysis: Actual vs. predicted plots were generated for both log-transformed and original price values. These graphs confirmed that the linear regression model effectively captured general pricing trends across most of the dataset.

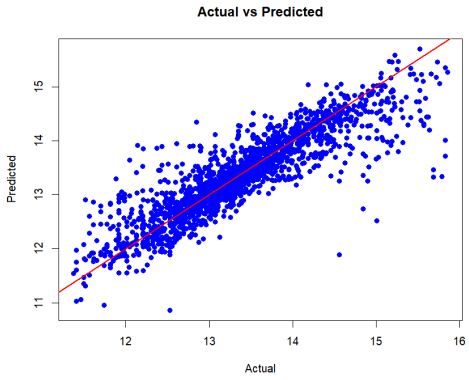

Fig. 1. Linear Regression for log price.

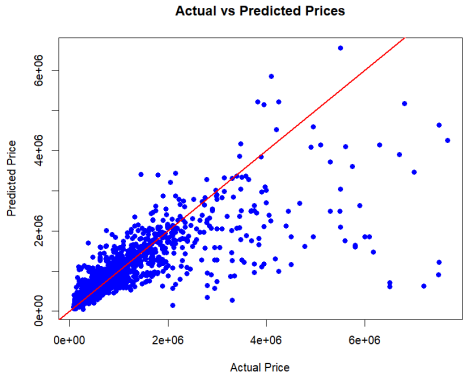

Fig. 2. Linear Regression for actual price.

C. Decision Tree Regression

The Decision Tree regression model was evaluated using RMSE and R-squared metrics to quantify prediction error and variance explained.

- **Model evaluation:** Initial R^2 of 69.3% improved to 73.5% after optimizing train-test split with a representative random seed.
- **Pruning results:** Pruning reduced accuracy, suggesting that an unpruned tree better captured relevant price patterns for this dataset.

The Decision Tree regression model was applied to capture non-linear relationships and feature interactions.

- **Model Choice:** Decision trees were selected due to their ability to:
  - Handle both numerical and categorical variables naturally
  - Require no assumptions about data distribution
  - Provide high interpretability for understanding feature influence
  - Serve as an exploratory model and benchmark for comparison with more advanced methods
- **Model Visualization:** The trained tree was visualized to examine:
  - Feature importance
  - Decision rules for price segmentation
  - Model complexity and structure

The visualization confirmed that age, km_driven, and premium brands were key determinants of price. Decision rules provided insight into pricing patterns and market segmentation.

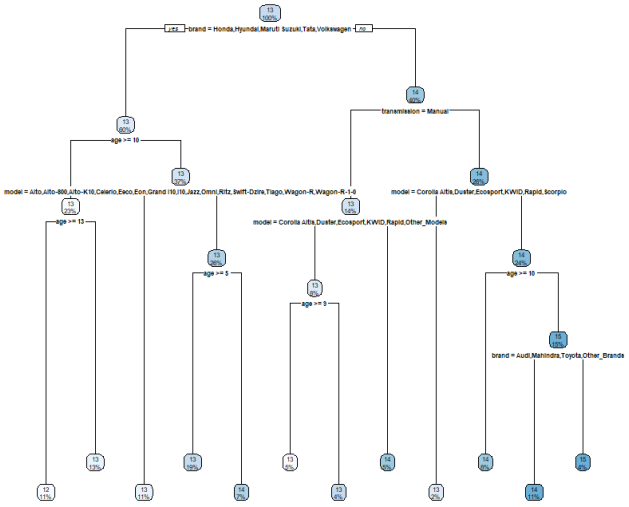

Fig. 3. No pruning decesion tree.

D. XGBoost Regression

The XGBoost model was evaluated using RMSE, MAE, and R^2 to measure predictive accuracy and generalization capability.

- **RMSE:** $\sim$0.33
- **MAE:** $\sim$222,000
- **R-squared:** $\sim$0.85

These results indicate high predictive accuracy, demonstrating the model's ability to generalize to unseen data while capturing complex interactions among features. XGBoost (Extreme Gradient Boosting) was used to improve predictive performance over single decision trees.

Mechanism:

1) **Initialization:** The model starts with a simple prediction, such as the mean of the target variable.
2) **Sequential Tree Building:** Trees are added iteratively, each predicting the residual errors of the current model.
3) **Gradient-Based Optimization:** Trees are constructed based on the gradient of the loss function rather than direct residuals.
4) **Regularization:** Overly complex trees are penalized using tree depth constraints and L1/L2 penalties on leaf weights.
5) **Learning Rate & Early Stopping:** Training stops when performance on validation data ceases to improve.

Advantages:

- Captures nonlinear relationships between car features and price
- Automatically models feature interactions, e.g., age $\times$ brand
- Strong regularization reduces overfitting
- Handles high-cardinality categorical variables
- Efficient for large structured datasets

Disadvantages:

- Less interpretable than linear regression
- Requires careful hyperparameter tuning
- Computationally heavier than simple models

Feature Importance and Insights: Using the importance function in XGBoost, km_driven and certain brands were identified as the leading factors affecting predicted prices.

Predicted vs. Actual Prices: A plot was generated to compare predicted versus actual prices, showing that most prices were concentrated below 2×10^6, where the model performed exceptionally well.

E. Random Forest Regression

The Random Forest regression model used 170 trees, $mtry = 7$, and $nodesize = 6$ to predict log-transformed car prices.

- **MAE:** 163,751
- **RMSE:** 322,288
- **R-squared:** 0.8555

Predictions closely followed the actual values, and residual analysis showed no significant bias. This confirms that the

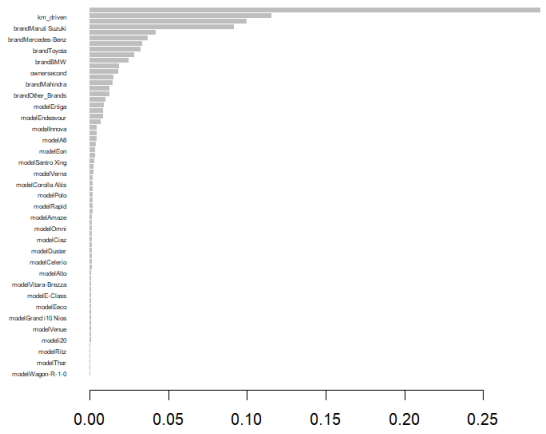

Fig. 4. Feature importance.

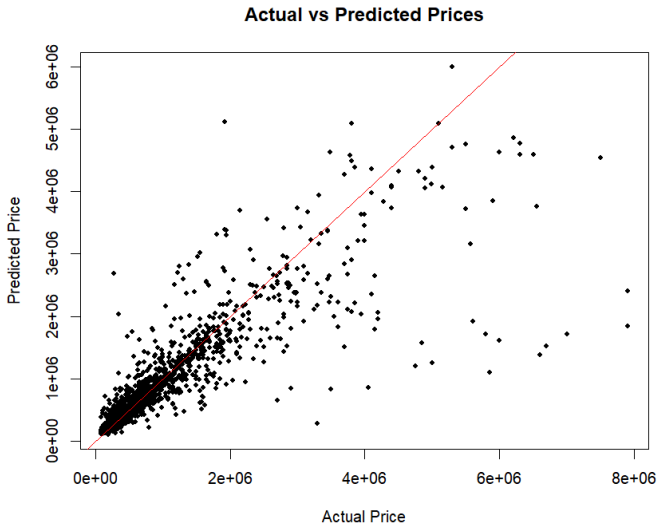

Fig. 5. Actual price VS Predicted price

Random Forest model effectively captures complex nonlinear relationships between vehicle attributes and price.

Random Forest Regression was applied as an ensemble method to enhance stability and capture complex interactions.

Importance of Seed Optimization (Model Stability Analysis): Iterating through multiple random seeds addresses the variability in model performance introduced by train-test splits:

- **Handling Randomness of Data Split:** The performance metric (e.g., R^2 or RMSE) is calculated on a single test set. A single split may include easier-to-predict points, giving an overly optimistic metric. Multiple seeds ensure results are not biased by a "lucky" split.
- **Measuring Model Stability:** Recording performance metrics for each seed allows computation of range, standard deviation, and average performance. A stable model shows a small range of R^2 values across seeds; a large range indicates dependency on the data split.

Impact of Random Forest Hyperparameters:

1) **mtry (Number of Variables Sampled at Each Split):**
   - *Increase mtry:* If mtry is close to the total number of features, trees become very similar. This increases correlation between individual trees, making the model more prone to overfitting (higher variance). Overall performance becomes similar to a single, deep decision tree.
   - *Decrease mtry:* If mtry is very small, the model has limited feature choices per split, leading to weaker individual trees and potential underfitting (higher bias). However, lower correlation reduces overall variance of the ensemble.
   - *Optimal mtry:* The goal is to minimize correlation between trees while ensuring individual trees remain strong predictors.
2) **ntree (Number of Trees to Grow):**
   - *Increasing ntree:* Generally improves accuracy by averaging predictions from more independent trees, stabilizing output and reducing variance. Beyond a certain point, additional trees yield negligible improvements.
   - *Decreasing ntree:* Results in fewer trees, unstable predictions, and higher variance.
3) **Node Size (Minimum Observations per Leaf):**
   - *Small node size:* Allows trees to grow fully, creating nodes highly specific to noise or outliers $\rightarrow$ high variance, risk of overfitting.
   - *Large node size:* Restricts tree growth, reduces variance, introduces some bias, but improves generalization.

Visualization and Residual Analysis:

- *Actual vs. Predicted Plot:* Confirmed predictions align closely with actual prices.
- *Residuals vs. Predictions Plot:* Errors were well-distributed with no significant bias, indicating robust model performance.

These analyses demonstrate that Random Forest effectively captures non-linear feature interactions while providing stability and generalization, making it suitable for reliable used-car price prediction.

V. Discussion

The results of this study demonstrate that ensemble-based models significantly outperform traditional regression and single-tree approaches in predicting used car prices. While Linear Regression effectively captured general pricing trends and provided interpretability, it was limited in handling nonlinear relationships. Decision Tree Regression improved flexibility but showed instability depending on data partitioning. The

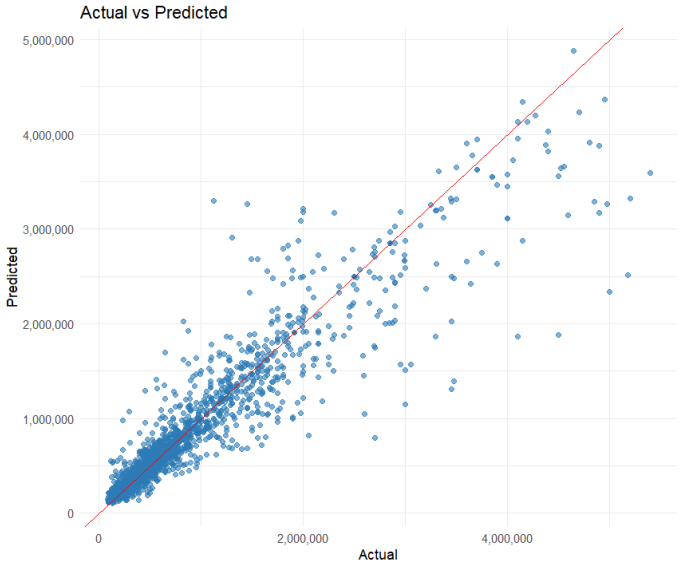

Fig. 6. Actual VS Predicted.

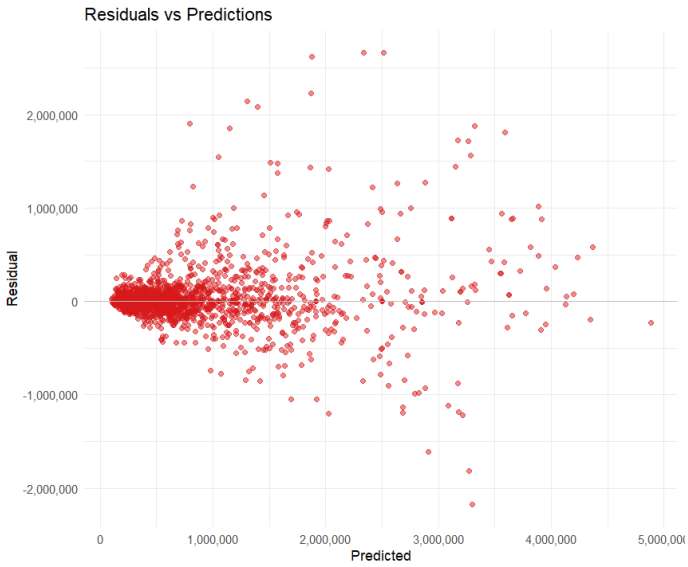

Fig. 7. Residuals VS Predictions.

primary payoff of this study lies in demonstrating that Random Forest and XGBoost offer superior predictive accuracy and robustness for real-world used car pricing tasks.

Several limitations were observed during experimentation. Model performance, particularly for Decision Trees, was sensitive to train–test splits, which was addressed through seed optimization and stability analysis. Additionally, high-cardinality categorical variables required careful preprocessing to avoid overfitting, which was mitigated through grouping and encoding strategies applied during preprocessing. No external data sources were introduced, limiting the scope to the available dataset.

From an interpretive perspective, the results confirm that vehicle age, kilometers driven, and brand-related attributes are dominant price determinants, consistent across all models. Ensemble methods were better able to capture nonlinear interactions among these features, leading to improved accuracy and generalization. Feature importance analysis in XGBoost and stability analysis in Random Forest further reinforced these findings.

The results are consistent with prior research indicating that ensemble learning methods outperform traditional regression models for structured tabular data. However, this study moves beyond previous work by providing a unified comparison of statistical, tree-based, and ensemble models under a consistent preprocessing and evaluation framework using a large, real-world Indian market dataset. This strengthens the validity of the findings and highlights the practical suitability of ensemble methods for used car price prediction.

VI. CONCLUSION

This study aimed to predict used car prices using a structured dataset of 9,238 vehicles and to evaluate the performance of multiple regression and ensemble learning models. The findings show that ensemble models significantly outperform traditional approaches. While Linear Regression and Decision Tree Regression were able to capture general pricing trends, XGBoost and Random Forest achieved the highest predictive accuracy. XGBoost obtained an R^2 of approximately 0.85, demonstrating strong generalization and effective modeling of nonlinear relationships, while the Random Forest model achieved the best overall performance with an R^2 of 0.8555, RMSE of 322,288, and MAE of 163,751, confirming its robustness and accuracy in price prediction. One important contribution of this paper is the comprehensive comparison of four predictive models on a real-world used car dataset, supported by detailed performance metrics, residual analysis, and stability evaluation. The results highlight the effectiveness of ensemble learning methods—particularly XGBoost and Random Forest—for structured tabular data and complex pricing problems. This work provides a valuable benchmark for future research and practical applications in automotive market analysis and price prediction systems.

REFERENCES

[1] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed. Wiley, 2012.
[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.
[3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.