INCITEST
International Conference on Informatics Engineering, Science & Technology
UNIKOM

Performance Evaluation of Hyperparameter-optimized Random Forest Regressor for Used Car Price Prediction.

Supriadi Supriadi [*], Irfan Dwiguna Sumitra

Master of Information Systems, Universitas Komputer Indonesia, Bandung, Indonesia

[*]Corresponding E-mail: supriadi75124007@mahasiswa.unikom.ac.id

Abstract. This study evaluates the performance of a hyperparameter-optimized random forest algorithm for predicting used car prices. The primary objective is to develop a robust model capable of delivering accurate price predictions. The random forest technique was selected for this study due to its proven effectiveness in handling non-linear regression problems. We utilized a public dataset from Kaggle, sourced via web scraping from Mobil123. The GridSearchCV method was employed to tune the hyperparameters and identify the optimal model configuration. The resulting model demonstrates strong predictive power, explaining over 93% of the price variance (R^2 score > 0.9349). Furthermore, the model's robustness is confirmed by an average cross-validation score of 0.9418. These results affirm that the optimized random forest model is a highly effective tool for this application. This research has practical implications for the automotive market, providing both buyers and sellers with a data-driven tool for more accurate price valuation.

Keywords: Price Prediction, Used Cars, Machine Learning, Random Forest Regressor, Hyperparameter Optimization.

1. Introduction

Predicting used car prices is becoming more and more necessary to close the gap between buyers and sellers as the number of private automobiles rises and the used car industry grows. It is easier for sellers to set a fair asking price when buyers are aware of a used car's resale value and can negotiate a price. A price that takes into account numerous factors, such as car make, year of manufacture, mileage, and condition, is challenging to find. One machine learning technique that can be applied to this problem is regression, which uses historical vehicle data to more accurately predict used car prices (Reddy and Kumar 2024)(Yadav, Kumar, and Yadav 2021). Predictive models that examine past data and generate more impartial pricing estimates and predictions can be created using machine learning techniques like the Random Forest algorithm (Solayman et al. 2023). As used car sales surge, dealers tend to charge unethical prices. To address this, we need a model that can calculate the best price using supervised learning (Alexstan et al. 2023).

The most crucial method for creating adaptable pricing prediction systems in a variety of industries, including real estate, transportation, and online shopping, is machine learning (Kalampokas et al. 2023). Machine learning is used to develop pricing algorithms that react swiftly to market conditions. In addition to maximizing service provider revenue, this price strategy is utilized to affect customer behavior and lessen the need for manual vehicle

repositioning (Teusch et al. 2023). Research from (Patanwala et al. 2023), implies that in order to narrow the gap between buyer demand and seller supply, a Random Forest-based regression model for precise price projection is essential. (Longani, Potharaju, and Deore 2021) According to their research findings, the XGBoost model outperformed Random Forest in predicting used car prices, with an RMSE of 0.53 that was significantly higher than Random Forest's RMSE of 3.44. Linear regression performs admirably, with a coefficient of determination (R) of 0.73. They looked into using linear regression to forecast the cost of secondhand cars (Muti and Yildiz 2023). Furthermore, research by (Prof. Dipti A. Gaikwad, Pratik S. Suwarnakar, Yash R. Mahajan 2023) RF, SVM, Lasso Regression, and Linear Regression are four machine learning techniques used to forecast used car prices. With an R2 value of 0.8697, the Random Forest technique outperformed the other strategies, suggesting that it is the best approach for price predicting (Patanwala et al. 2023). It attempts to integrate several decision trees to increase accuracy by averaging the outcomes from each one (Aszhari et al. 2020).

According to study (Ghule, Kim, and Jang 2023), Although a default model might work well, the author pointed out that hyperparameter tuning approaches are necessary to achieve better outcomes, such as a greater R2 (R-squared) value and a lower MARE (Mean Absolute Relative Error). Research by (Warkentin et al. 2022), shows how to effectively choose hyperparameters by first identifying the best hyperparameters using grid search, then conducting a random search in a more targeted area surrounding the set of hyperparameters that perform the best.

This study is to evaluate the performance of the random forest algorithm with hyperparameter optimization for predicting used car prices and to create a robust model that can make predictions and provide results to help consumers make informed choices about used carsprovide results to help consumers make informed choices about used cars.

2. Method

2.1 Research Flow.

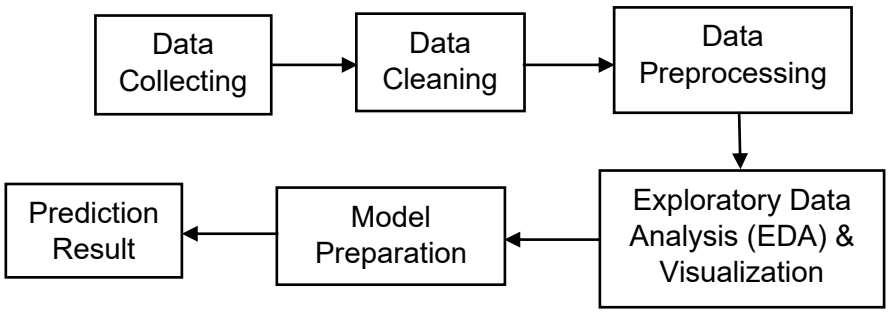

Figure 1. Architecture diagram.

As seen in Figure 1, this research flow takes a methodical approach with multiple consecutive steps. To guarantee the legitimacy of the data and its preparedness for analysis, the procedure starts with data collection, cleaning, and pre-processing. Exploratory Data Analysis (EDA), the following step, uses visualization to obtain deeper insights and comprehend the properties of the data. The insights gained from EDA then serve as a fundamental reference in the model preparation stage, where the predictive model architecture is designed and configured. The final stage of this workflow is model implementation to produce final predictions results.

2.2 Data Collection

The dataset for this study comes from the publicly available Kaggle platform using web scraping from Mobil123. Attributes include price, make, model, and year of manufacture, all of

which can influence market prices. The dataset on this platform was chosen because it best aligns with the study's objective, which is to test the random forest algorithm with hyperparameter optimization to create a predictive model.

2.3 Data Cleaning

The data cleaning phase is a crucial pre-modeling process to ensure dataset quality. This process removes any data that is missing, inconsistent, or superfluous and could have an impact on the prediction model's performance. The "seller" field is removed, as are other columns that have a high percentage of missing values or do not aid in the prediction process. Rows with missing data are also eliminated during the column deletion process. Furthermore, records that have values in the "condition" column but do not fall into any of the identified categories are filtered out. The analysis and modeling stage will only employ relevant, consistent, and clean data, a thorough data cleaning process will yield more precise prediction results.

2.4 Data Preprocessing

Prior to analysis and predictive modeling, the cleaned data is subjected to additional processing in the data pretreatment step. The goal of this stage is to transform the raw data into a machine-readable format. In preprocessing, we convert several categorical attributes, including make, model, transmission, condition, fuel type, body type, and mileage, into numeric values so that the machine can understand the data we label. This is achieved by using label encoding, which assigns a numeric representation to each category contained in a variable. Next, we modify character data types, such as engine volume (ENGV), price, and year, to ensure the data format is suitable for the operations and processing performed by the algorithm. This preprocessing step is crucial for maintaining the consistency of the data structure.

2.5 Exploratory Data Analysis (EDA) & Visualization

The exploratory data analysis (EDA) can be used to get a preliminary picture of the distribution of the data, identify outliers, and determine how the variables connect to the modeling goals.

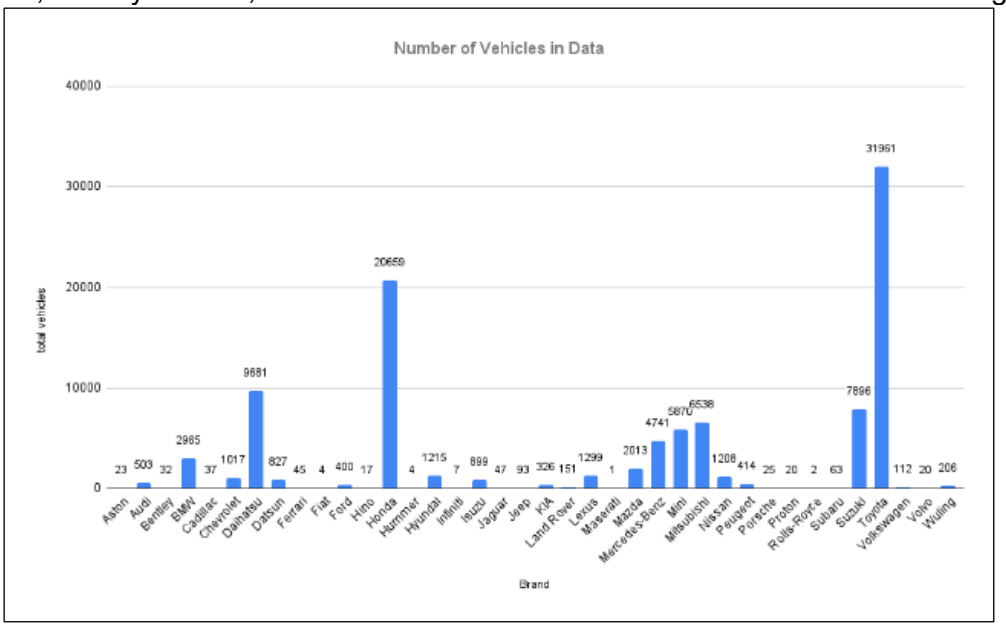

Figure 2. Bar Chart.

In order to generate a numerical summary of every attribute, including Total Vehicles and Brand, this analysis involves the development of descriptive statistics. Several visualization approaches are employed to facilitate the interpretation of data, such as bar charts (Figure 2)

that show the distribution of vehicle numbers by category (brand, total cars). With 31,950 units, Toyota is the brand with the most automobiles, followed by Honda (20,469 units) and Suzuki (7,886), as seen in Figure 2. Nissan, Mitsubishi, and Daihatsu are among the other brands that significantly contribute to the data composition.

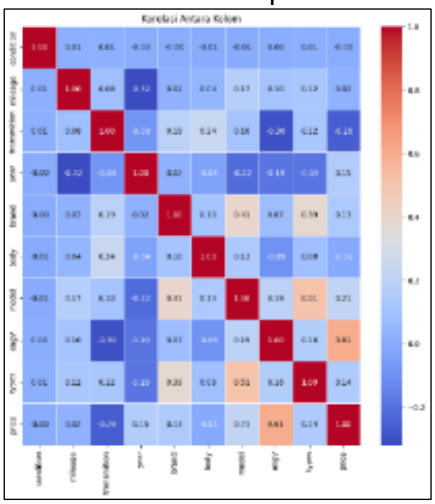

Figure 3. Correlation between columns.

As shown in Figure 3, with a value of 0.61, engV/engine volume is the variable most strongly correlated with car price, suggesting that engine capacity influences pricing. While transmission has a negative correlation of -0.26, suggesting that transmission influences price, the model, year, and brand variables exhibit extremely small positive correlations with price (0.21, 0.15, and 0.13, respectively). In contrast, other characteristics like body, mileage, and condition show very little correlation or are almost zero, meaning they have no bearing on pricing. There is a correlation between the car's shape and its kind, as evidenced by the separation of correlations between variables, such as types with a model of 0.51 and models with types of 0.51. This condition suggests that mechanical parameters, including engine volume, have a bigger impact on price than the car's overall condition or features.

2.6 Model Preparation

A group of columns from the dataframe df, including "types," "engV," "model," "brand," and "year," which are known to have an impact on the selling price of used automobiles and are categorized as predictor variables, are extracted to create the feature variable (X) during the Model Preparation stage. In the meantime, the attribute that the model would forecast, the "price" column, is used to formulate the target variable (y). This feature selection is crucial because it filters characteristics that lower model complexity, which affects training efficiency and accuracy. The train_test_split function from the sklearn library is then used to partition the dataset (dataset partitioning) into two datasets: the training set and the testing set. 20% of the total observations will be utilized as test data (x_test, y_test), with the remaining 80% being used as training data (x_train, y_train), according to the parameter test_size=0.2. The random_state=42 option is set to ensure that the data splitting process produces reliable findings and allows the experiment to be repeated.

2.7 Prediction Result

The Random Forest Regressor model is used through a pipeline. The purpose of this pipeline is to make it easier to combine data transformation with model training. To improve model performance, hyperparameters are optimized using the GridSearchCV approach. The following are the best parameter outcomes from the tuning process: "regressor_max_depth": 20, "regressor_min_samples_split": 5, and "regressor_n_estimators": 200 are the optimal parameters. Maximum Points: 0.9418338432635748 After training, the optimal model

produces an R-squared (R^2) value of 0.94 on the training data, meaning that the model can account for 94% of the variation in the target data. With the use of vehicle details like year, mileage, engine capacity, this model may be used to predict the cost of used cars. It has ideal specifications and a high R2 value. The metrics R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are used to assess the model's performance on the test data once the optimal model has been obtained through the hyperparameter optimization process using GridSearchCV. This step aims to assess the model's generalizability to data that hasn't been examined previously and compare the performance results after adjusting the hyperparameter to the default parameter.

3. Results and Discussion

3.1 Results

In the final stage of this research, the trained prediction model is deployed to estimate used car prices and evaluate its performance. The model, a RandomForestRegressor, was developed within a pipeline using the optimal hyperparameters previously identified by GridSearchCV. This fully trained model predicts the target variable (price) based on key features, including type, engV, model, brand, and year.

As an example, the manual input data provided is [[0, 1.5, 2, 3, 2022]], which means a combination of encoded specifications (types, engV, model, and brand) and the year of production. The predicted result of the model for this input is an estimated price of Rp254,421,526.11. This estimate is obtained from the average of all decision trees (sklearn.base.BaseEstimator - RandomForestRegression) in the Random Forest ensemble, which takes into account the existing relationship patterns between these input features and the car prices studied during the training process.

3.1.1 Performance Evaluation on Testing Data

The performance evaluation demonstrates the effectiveness of the hyperparameter-optimized RandomForestRegressor model. It achieved an R^2 score of 0.9349, indicating that the model can explain 93.49% of the variance in used car prices. The model's Mean Absolute Error (MAE) of Rp21,528,621.29 signifies that, on average, its predictions deviate from the actual price by approximately Rp21.53 million. Furthermore, the Root Mean Squared Error (RMSE) was Rp99,076,620.27. Since RMSE penalizes larger errors more heavily than MAE, this substantially higher value suggests that while the model is generally accurate, it produces a few predictions with significantly larger errors.

The following table (Table 1) displays the results of the model evaluation, including performance metrics on the test data and the best results from the GridSearchCV cross-validation process:

Table 1. Model Evaluation Results.

Metrics	Value
R-squared (R^2)	0.9349
Best Score (CV R^2)	0.9418
Best Hyperparameters	{‘regressor_max_depth’:20,’regressor_min_samples_split’:5 ,’regressor_n_etimator’:200}

The results of the performance evaluation are shown in Table 1. The model's ability to explain prices is demonstrated by the R-squared (R^2) value of 0.9349, while the Best Value (CV R^2) is 0.9418. The model is robust and does not overfit, as evidenced by the closeness of the two values. A max_depth of 20, min_samples_split of 5, and n_estimators of 200 were the best hyperparameter combinations discovered during the tuning process and utilized to attain optimal performance. These results imply that changing hyperparameters can significantly improve the prediction accuracy of the random forest model.

In addition to the numerical evaluation, several visualizations were created to provide a clearer picture of the model's performance:

3.1.2 Feature Importance Chart

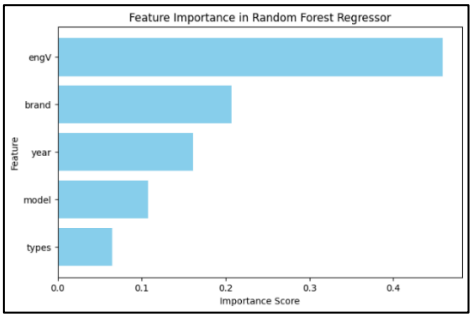

Figure 4. Feature Importance Chart

The degree to which each attribute influences used car price estimates is depicted in Figure 4. The biggest factor influencing used car costs is engine volume (engV), which is followed by brand and year. In contrast, type and model have less of an effect on the outcomes of predictions. This data can help determine which factors are most important and serve as a foundation for feature selection in subsequent studies.

3.1.3 Actual vs. Predicted Scatter Plot

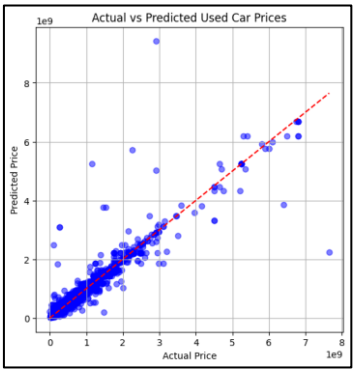

Figure 5. *Actual vs. Predicted Scatter Plot*

Figure 5, shows the difference between the model's predicted and real used automobile pricing. Points along the diagonal red line show how well the model's predictions match the real prices. With comparatively minimal mistakes, the majority of the points are grouped around the diagonal line, suggesting strong predictive ability. Cases where the predictions were substantially off are indicated by points far from the line, these are probably the result of outliers or special data requirements.

3.2 Discussion

The results demonstrate the validity and practical utility of the optimized Random Forest Regressor model in the complex task of used car price prediction. The chosen features type, engine volume, model, brand, and year together account for almost 93.5% of the price

variability, as indicated by the high R-squared value of 0.9349 on the test data, demonstrating their great predictive potential.

A deeper analysis of the error metrics provides crucial insights into the model's real-world performance. The model's price estimate is, on average, off by this amount, according to the MAE of Rp 21.53 million. This degree of error can be regarded as realistically acceptable for general value in the Indonesian used car market. The much greater RMSE of Rp 99.08 million, however, shows that although the model is often accurate, it occasionally generates forecasts with huge mistakes. This difference between MAE and RMSE is typical of models that deal with outliers, it suggests that certain vehicles perhaps rare models, heavily modified cars, or those with unusual conditions not captured by the existing features are priced incorrectly by a large margin. In contrast to the study (Alhakamy et al. 2023), projected used car prices using a linear regression model and shown how prices varied according to a number of parameters, indicating that more research using more straightforward yet efficient techniques can still provide valuable insights into the factors influencing price, and research by (Dutulescu et al. 2023), shows how more accurate pricing predictions may be obtained by using deep learning algorithms to find significant elements impacting auto market prices.

Although each strategy has its own pros and cons, combining various ways can result in a more thorough understanding of the dynamics of the automobile industry. The primary limitation of the current model is its reliance on a limited feature set. Key price-determining factors such as mileage (odometer reading), vehicle condition, transmission type, service history, and geographical location were not included. The absence of these features is a likely explanation for the large errors reflected in the high RMSE.

4. Conclusion

This research successfully created and validated an improved Random Forest Regressor model. With an R-squared value of 0.9349, a Mean Absolute Error of Rp 21,528,621.29, and a Root Mean Squared Error of Rp 99,076,620.27, the final model demonstrated excellent generalization performance on the test data. During training, the model's best cross-validation R-squared score was 0.9418, demonstrating its robustness.

These findings show that the model can produce reliable price forecasts and correctly explain over 93% of the variation in used car prices. Additionally, the analysis confirms that hyperparameter optimization via GridSearchCV significantly enhanced model performance compared to the default configuration. The synergy of machine learning, the Random Forest algorithm, and rigorous optimization proves to be a highly effective methodology for this prediction task.

For future research, several enhancements are recommended to improve the accuracy and generalization of the model, including descriptive variables such as mileage, vehicle condition, transmission type (automatic/manual), service history, and color, which provide important context for the assessment. In addition to expanding the feature set, future research should explore and compare the performance of other advanced machine learning algorithms, such as XGBoost and Gradient Boosting, or more complex models such as Artificial Neural Networks, which may capture underlying data patterns differently, which will guide future improvements in both feature engineering and model selection.

References

Alexstan, Aarone Steve J, Krishna M Monesh, M Poonkodi, and Vineet Raj. 2023. “Used Car Price Prediction Using Machine Learning.” *Advances in Science and Technology* 124: 512–17. doi:10.4028/p-9x4ue8.

Alhakamy, A’aeshah, Areej Alhowaity, Anwar Abdullah Alatawi, and Hadeel Alsaadi. 2023. “Are Used Cars More Sustainable? Price Prediction Based on Linear Regression.” *Sustainability* 15(2): 911. doi:10.3390/su15020911.

Aszhari, F. R., Z. Rustam, F. Subroto, and A. S. Semendawai. 2020. "Classification of Thalassemia Data Using Random Forest Algorithm." *Journal of Physics: Conference Series* 1490(1). doi:10.1088/1742-6596/1490/1/012050.

Dutulescu, Andreea, Andy Catruna, Stefan Ruseti, Denis Iorga, Vladimir Ghita, Laurentiu Marian Neagu, and Mihai Dascalu. 2023. "Car Price Quotes Driven by Data-Comprehensive Predictions Grounded in Deep Learning Techniques." *Electronics (Switzerland)* 12(14): 1–25. doi:10.3390/electronics12143083.

Ghule, Balaji G, Min-Kyeong Kim, and Ji-Hyun Jang. 2023. "Predicting Photoresist Sensitivity Using Machine Learning." *Bulletin of the Korean Chemical Society* 44(11): 900–910. doi:10.1002/bkcs.12776.

Kalampokas, Theofanis, Konstantinos Tziridis, Nikolaos Kalampokas, Alexandros Nikolaou, Eleni Vrochidou, and George A. Papakostas. 2023. "A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques." *IEEE Access* 11(March): 46627–43. doi:10.1109/ACCESS.2023.3274669.

Longani, Chetna, Sai Prasad Potharaju, and Sandhya Deore. 2021. "Price Prediction for Pre-Owned Cars Using Ensemble Machine Learning Techniques." doi:10.3233/apc210194.

Muti, Sümeyra, and Kazım Yildiz. 2023. "Using Linear Regression For Used Car Price Prediction." *International Journal of Computational and Experimental Science and Engineering* 9(1): 11–16. doi:10.22399/ijcesen.1070505.

Patanwala, Adnan, Huzefa Polaiwala, Qusai Jetpurwala, Burhan Rampurawala, Swati Nadkarni, and Theres Bemila. 2023. "Prediction of the Value of Pre-Owned Cars Using Machine Learning." In *2023 6th International Conference on Advances in Science and Technology (ICAST)*, , 209–12. doi:10.1109/ICAST59062.2023.10455055.

Prof. Dipti A. Gaikwad, Pratik S. Suwarnakar, Yash R. Mahajan, Amita. 2023. "Used Car Price Prediction Using Random Forest Algorithm." *International Journal For Multidisciplinary Research* 5(3): 1–9. doi:10.36948/ijfmr.2023.v05i03.3308.

Reddy, S. Naveen, and S. Kumar. 2024. "A Comparative Analysis of XGBoost Model and AdaBoost Regressor for Prediction of Used Car Price." (AI4IoT 2023): 441–46. doi:10.5220/0012510700003739.

Solayman, Sanzida, Sk. Azmiara Aumi, Chand Sultana Mery, Muktadir Mubassir, and Riasat Khan. 2023. "Automatic COVID-19 Prediction Using Explainable Machine Learning Techniques." *International Journal of Cognitive Computing in Engineering* 4: 36–46. doi:10.1016/j.ijcce.2023.01.003.

Teusch, Julian, Jan Niklas Gremmel, Christian Koetsier, Fatema Tuj Johora, Monika Sester, David M. Woisetschlager, and Jorg P. Muller. 2023. "A Systematic Literature Review on Machine Learning in Shared Mobility." *IEEE Open Journal of Intelligent Transportation Systems* 4(November): 870–99. doi:10.1109/OJITS.2023.3334393.

Warkentin, Matthew T, Hamad Al-Sawaihey, Stephen Lam, Geoffrey Liu, Brenda Diergaarde, Jian-Min Yuan, David O Wilson, et al. 2022. "Radiomics Analysis to Predict Pulmonary Nodule Malignancy Using Machine Learning Approaches." doi:10.1101/2022.10.03.22280659.

Yadav, Anu, Ela Kumar, and Piyush Kumar Yadav. 2021. "Object Detection and Used Car Price Predicting Analysis System (UCPAS) Using Machine Learning Technique." *Linguistics and Culture Review* 5(S2): 1131–47. doi:10.21744/lingcure.v5ns2.1660.