Stacked Generalization with Sequential-Model based Optimization for estimating Used Car Valuation in Indonesia

Isti Surjandari
Industrial Engineering Department, Faculty of Engineering, Universitas Indonesia, Indonesia
isti@ie.ui.ac.id (corresponding author)

Ahmad Dzikri
Industrial Engineering Department, Faculty of Engineering, Universitas Indonesia, Indonesia
ahmad.dzikri@ui.ac.id

Arian Dhini
Industrial Engineering Department, Faculty of Engineering, Universitas Indonesia, Indonesia
arian@ie.ui.ac.id

Enrico Laoh
School of Industrial Engineering and Management, Oklahoma State University, Stillwater, USA
elaoh@okstate.edu

Kinanthy D. Pangesty
Industrial Engineering Department, Faculty of Engineering, Universitas Indonesia, Indonesia
kinanthy.dwi@ui.ac.id

Pocut S. Aurora
Industrial Engineering Department, Faculty of Engineering, Universitas Indonesia, Indonesia
pocut.shafira@ui.ac.id

Dewa Ferrouzi
Industrial Engineering Department, Faculty of Engineering, Universitas Indonesia, Indonesia
dewa.ferrouzi@ui.ac.id

Received: 26 June 2024 | Revised: 9 August 2024 | Accepted: 24 August 2024

 DOI: https://doi.org/10.48084/etasr.8226

ABSTRACT

In Indonesia, the purchase and sale of used vehicles is a common practice. The valuation of a used vehicle is influenced by several factors, making it challenging to determine an appropriate selling price. To address this issue, this study employs a stacked generalization (stacking) algorithm to integrate Machine Learning (ML) techniques that have demonstrated efficacy in prior research on used car valuations. The Sequential Model-Based Optimization (SMBO) algorithm is employed to achieve high accuracy while ensuring an efficient hyperparameter optimization process. The initial price of a vehicle is undoubtedly a significant determinant of its resale value. However, this fact is frequently overlooked in previous studies on developing car price estimation models. This study makes a contribution to the field by addressing this issue. The use of the initial price as an input for the model enables two distinct types of analysis: one for the assessment of used car prices and the other for the measurement of the degree of residual valuation of used cars in relation to their initial costs. The results demonstrated that the optimized stacking model exhibited superior predictive ability compared to the other algorithms in both analyses. Feature analysis substantiated the considerable influence of the initial price on the used car's price. This study also

corroborates the assertion that accurately predicting the valuation of a used car cannot be achieved by solely considering the usage of the previous owner, such as the car's age and mileage. It is crucial to take into account the car's original attributes, particularly its initial price.

Keywords-used car valuation; residual value; stacked generalization; sequential model-based optimization; feature analysis

I. INTRODUCTION

The transportation sector plays a critical role in the economic and social development of any country. In developing countries where public transportation is limited and population density is high, this often results in an increase in private car ownership [1, 2]. As reported by Statistics Indonesia (BPS), the number of private cars in Indonesia has increased since 2015 [3]. Notwithstanding the adverse impact of the SARS-CoV-2 pandemic on Indonesia's trade sector, including new car sales, there has been a notable surge in interest in the used car market, as reported by the Association of Indonesia Automotive Industries (Gaikindo) in 2020 [4]. This is corroborated by the fact that 55.7% of Indonesians expressed a preference for purchasing a used car during an economic downturn [5]. Used cars are also prevalent in other developing countries due to their affordability [6]. As a country with a large population, these figures indicate significant potential in the used car market in Indonesia. The valuation of a used car is influenced by a number of factors, including the brand, year of manufacture, and mileage of the vehicle [5, 7]. This results in a range of potential selling prices. An accurate pricing model for used cars can assist a number of parties. Firstly, it can help sellers determine the selling price of used cars rationally and quickly. Secondly, it can help prospective buyers estimate a reasonable purchase price for used cars and compare several cars simultaneously. Thirdly, it can help leasing companies estimate the residual value of cars after the lease term ends [8].

Prior studies have concentrated on the development of car valuation models employing ML methodologies, encompassing a range of approaches, from classical techniques [7, 9-10] to ensemble methodologies [6, 11-13]. Authors in [9], evaluated the efficacy of the Adaptive Neuro-Fuzzy Inference System (ANFIS) algorithm, in comparison to that of Neural Networks (NN) for the purpose of modeling used car prices. Meanwhile, authors in [11] conducted a comparative analysis of the performance of 19 distinct ML algorithms. The ensemble method yielded a more precise estimation model, with the Ensemble Selection (ES) and Random Forest (RF) algorithms emerging as the most promising. This finding was consistent with the results of another study, in which the RF algorithm, optimized through random search, demonstrated high estimation accuracy. Authors in [12] proposed an ensemble-based method that combined NNs and the support vector algorithm to estimate used car prices, resulting in high accuracy. In contrast, authors in [10] found that the Support Vector Regression (SVR) algorithm performed better after optimization through feature selection based on the Boruta algorithm. Furthermore, a previous study [7], has also demonstrated the beneficial impact of optimization on the performance of the NN algorithm-based model in estimating used car prices. Authors in [6], examined five regression models for their ability to predict used car prices in developing countries. The findings of the study indicated that the Gradient Boosting (GB)-based algorithm proved to be the most effective in terms of performance. In general, previous studies have demonstrated that each ML algorithm has a distinct level of success in estimating used car prices.

Previous studies have indicated that certain features are more influential in determining used car prices than others. These include car type, mileage, and age, as well as the vehicle model [7, 10, 12, 13]. It is a categorical feature that characterizes a specific car series (e.g., the Toyota Corolla). However, this feature often comprises an excess of distinct values, which can give rise to complications in the encoding process of ML modelling [14]. To address this challenge, an alternative approach has been employed, namely feature removal [13]. However, the removal of a feature necessarily entails a reduction in the information available for modeling, and thus the study replaced the car's model with the initial price. Theoretically, the initial price of an asset is correlated with its residual value [15]. Furthermore, the initial price is a numeric input feature that is preferred for most ML algorithms [16]. The ratio between the initial and selling prices of used cars has been employed as a dependent variable in previous studies, as referenced in [11]. This approach, however, may lead to confusion, as the model predicts the ratio rather than the actual selling price. While data regarding the initial price of used cars is available, it is employed only as a divisor for the current value of the vehicle in question, thus producing the dependent variable (the ratio).

A review of recent studies on estimating used car valuations reveals a notable absence, which is that the combination of the stacking algorithm with Sequential Model-Based Optimization (SMBO) has not been leveraged. However, both methods are regarded as promising due to their favorable outcomes in other studies. Authors in [17], concluded that the stacking algorithm demonstrated superior performance compared to a single ML method or a combination of similar ML methods. This is because integrating different ML methods in stacking allows for the optimization of the complementary strengths of distinct models. The stacking model is demonstrated to offer an accurate and resilient model that is adaptable to changing datasets [18]. Conversely, authors in [19] asserted that SMBO offers advantages in terms of efficiency when compared to the frequently used optimization method, namely the grid search method. In conclusion, the available evidence suggests that SMBO may offer advantages in terms of efficiency when compared to other optimization methods. The principal contributions of this study can be encompassed in the following:

- The objective of this study is to combine the stacking algorithm with SMBO in order to predict used car prices and residual valuation ratios with greater accuracy.

- A separate model was created to predict the selling price and remaining valuation ratio of used cars, with the aim of reaffirming the findings of previous studies on used car valuation.
- Using authentic data from the Indonesian used car market to guarantee the applicability of the findings for pricing determination in the Indonesian used car market.

II. METHODOLOGY

This section presents a description of the stacking algorithm and SMBO, along with a brief overview of the learning algorithms utilized in their development. The learning algorithms used in this study include RF, GB, SVR, and NN. Subsequently, the feature importance of the permutation technique employed in the analysis stage was elucidated.

A. Stacking

The technique of stacking was employed in numerous studies and has demonstrated encouraging outcomes in a number of fields, including short-term load forecasting [20], health diagnosis [21], and phishing attack detection [22]. The stacking algorithm represents a method of combining several lower-level algorithms to form a higher-level model with enhanced prediction accuracy. In the context of stacking algorithms, the term "base learner" is often used to refer to a lower-level algorithm, whereas the term "meta-learner" is used to describe a higher-level algorithm that is employed to generalize the base learner [23]. A variety of algorithms, including RF, GB, SVR, and NN, were frequently employed as base learners in stacking algorithms, with some demonstrating promising outcomes. RF is an ensemble learning method that modifies the bagging technique. In the context of regression problems, the predicted results of each tree in the RF model are averaged to determine the final prediction value [24]. The GB regression tree algorithm, henceforth referred to as GB, is a representative algorithm that employs boosting techniques. This algorithm improves the performance of the model gradually by adding decision trees in order to minimize a specific loss function [25]. SVR is a method for solving regression problems that functions in a manner analogous to that of SVM. The fundamental concept underlying SVR is the identification of the optimal linear fit, which is the line that encompasses the greatest number of data points [10]. The Multilayer Perceptron (MLP), one of the simplest forms of NN, is based on generalizations of linear models that undergo multiple processing stages before reaching a decision. In this study, the term NN is used to refer to the MLP algorithm.

B. Sequential Model-based Optimization

SMBO offers an effective methodology for identifying the global optimum value in problems where objective functions are challenging to evaluate and lack a defined structure (i.e., a "black box"). It is comprised of two principal elements: a surrogate model for estimating the objective function and an acquisition function for identifying the most promising point for evaluation while maintaining a balance between exploitation and exploration [26]. As a principal component of SMBO, the surrogate model is employed to estimate the response function of a dataset for specified hyperparameter settings, thereby facilitating the sequential evaluation of promising combinations [19]. As stated by authors in [26], the Gaussian Process (GP), is a distribution function which is determined by the average function $\mu(x)$ and the covariance function $k(x,x')$, is a type of surrogate model that has typically been selected in previous studies that incorporate SMBO due to its simplicity. The expressions for GP can be written as:

$$f(x) \sim GP\big(\mu(x);\ k(x,x')\big) \quad (1)$$

The acquisition function is a tool that assists in identifying the optimal point, which is defined as a point that potentially has the optimal value based on the objective function [26]. Expected Improvement (EI) is a type of acquisition function that has gained considerable popularity due to its intuitive nature and its ability to prevent the optimization process from becoming trapped in a local optimum solution [9].

C. Permutation Feature Importance

Permutation feature importance is a technique employed to evaluate the impact of each feature on the ML model. This is achieved by calculating the reduction in model performance when a feature is randomized. The process of randomization will result in the elimination of any correlation that may exist between the features and the model output. The feature with the highest permutation importance score will then be identified as the most important. The permutation importance score of a feature (imp_j) is calculated by subtracting the initial performance of the model from the average model performance after K randomizations. In other words, the permutation importance score of a feature is obtained by reducing the initial performance of the model by the average model performance after K randomizations, where every k randomization for feature j is symbolized by $s_{k,j}$ and is:

$$imp_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \quad (2)$$

III. EXPERIMENTAL DESIGN

This section outlines the pre-processing steps and elucidates the methodology employed in the construction of the proposed model, which is based on the stacking and SMBO algorithms. Figure 1 presents a flowchart of the research process.

A. Data Acquisition and Preprocessing

The dataset provides a comprehensive overview of the used car market in Indonesia at a specific point in time, encompassing a diverse range of brands, models, and conditions [27]. A more detailed examination of the data, is presented in Table I, with a total of 1,526 observations being collected and subjected to the following pre-processing steps:

- The data were subjected to cleaning procedures in order to eliminate superfluous features.
- Scraping was conducted four times in total, on each occasion for a different car body type: hatchback, Multi-Purpose Vehicle (MPV), sedan, and Sports Utility Vehicle (SUV). Accordingly, it was imperative to consolidate these four datasets.
- The data was processed to create new features with a standard format. For example, age and warranty expired

were changed from a text format to a date format. This meant calculating the difference between the current month and the date the data was retrieved. The warranty expired feature was converted to "No" to indicate that the car was no longer under warranty. Features related to the car's condition, such as exterior score, interior score, road test score, and underside score, were converted to ensure values ranged between 0 and 1, with a higher value indicating less damage.

- Categorical features with two values are binary-encoded (0 and 1). For features with more than two values, a dummy variable is generated using one-hot encoding [28].
- Feature scaling was performed to make each feature range in value from 0 to 1.

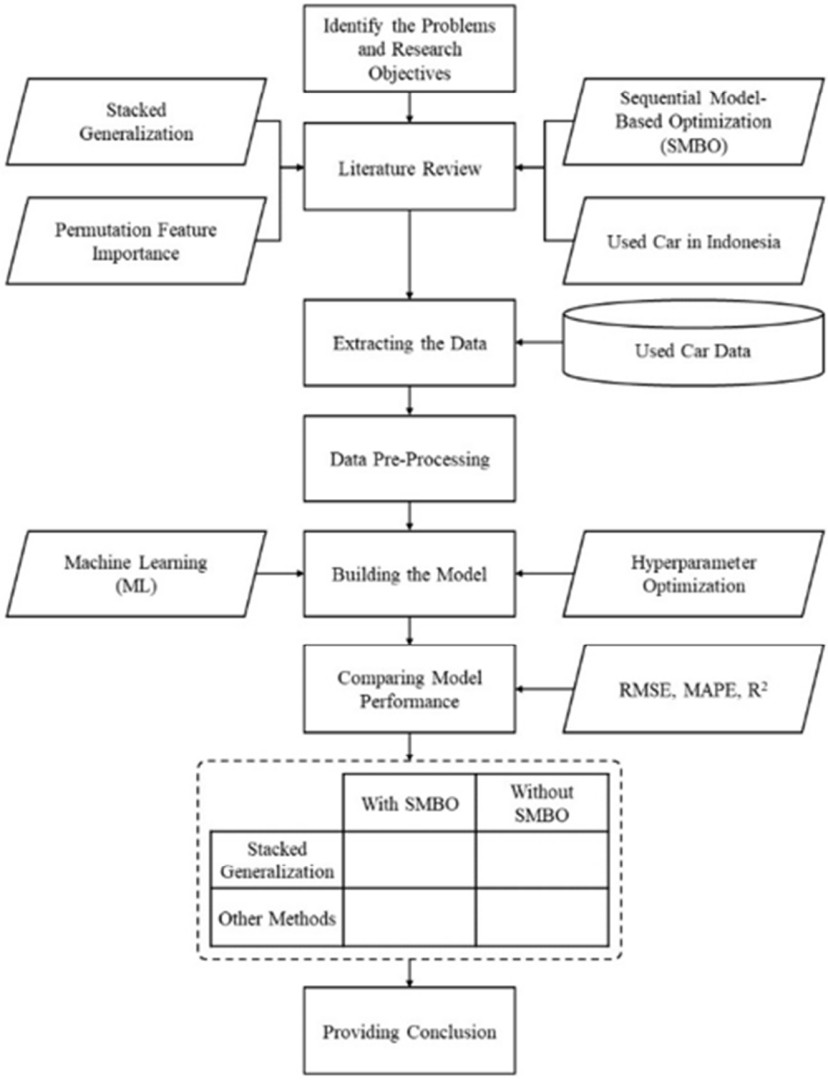

Fig. 1. Flowchart of the research process.

The data pre-processing steps resulted in nine categorical and nine numerical features. Tables I and II present the definitions and statistical data for the categorical and numerical features, respectively. Following pre-processing, the dataset was prepared for training with the ML algorithm. In accordance with the approach of other studies related to ML, the dataset was divided into two parts: 70% (1,068 observations) were used as a training dataset, and 30% (458 observations) were used as a testing dataset.

B. Modeling

This section presents an analysis of the two models that were developed: the used car price model and the residual valuation ratio model. The used car price model was constructed using the stacking algorithm and SMBO, and all the variables considered when estimating a car's selling price were incorporated. The stacking algorithm comprised the RF, GB, SVR, and NN algorithms as the base learners and Linear Regression (LR) as the meta-learner. The base-learner

algorithms were selected for their demonstrated efficacy in prior studies, while LR was chosen for its simplicity [29]. In this study, SMBO was performed for 50 iterations with GP as the surrogate model and EI as the acquisition function. GP was selected as the surrogate model due to its prevalence in the literature [30], while EI was chosen as the acquisition function because it demonstrated superior performance compared to other functions [31].

TABLE I. DESCRIPTION OF CATEGORICAL FEATURES

Name	Definition	Class (%)	
Brand	Manufacturing brand	Brand A (35.57)	Brand G (1.19)
		Brand B (33.33)	Brand H (0.89)
		Brand C (10.56)	Brand I (0.45)
		Brand D (8.93)	Brand J (0.30)
		Brand E (4.76)	Brand K (0.30)
		Brand F (3.57)	Brand L (0.15)
Car type	Car body style	MPV (41.07)	SUV (21.58)
		Hatchback (32.29)	Sedan (5.06)
Color	Dominant color on the car body	Black (27.83)	Silver (15.33)
		White (21.28)	Red (8.33)
		Grey (20.83)	Others (6.40)
Transmission	Gearbox type	Automatic (79.32)	Manual (20.68)
Fuel type	Recommended fuel for the car	Gasoline (92.41)	Diesel (7.59)
Seller_type	Previous owner	Individual (90.03)	Company (9.97)
Spare_key	Availability of spare key	Yes (79.61)	No (20.39)
Service book	Availability of service	Yes (80.80)	No (19.20)
N_of_seats	Number of seats	≥7 (53.87)	4 (3.72)
			5 (42.41)

TABLE II. DESCRIPTION OF NUMERICAL FEATURES

Feature	Definition	Mean	Std Dev
Initial_price	Price of the car in new condition (in 1000 IDR)	271,340	138,518
Mileage	Distance traveled by the car as shown on the speedometer (in km)	41,800	26,623
Engine_capacity	Car engine cylinder volume (cc)	1,514	355
Age	Age of the car (in months)	45.5685	18.5034
Warranty	Remaining car warranty (in months); 0 means that the warranty is over	2.3839	5.7360
Exterior_score	Car exterior condition; a score closer to 1 means that less damage is visible	0.8997	0.0711
Interior_score	Car interior condition; a score closer to 1 means that less damage is detected	0.9854	0.0199
Road_test_score	Test drive result; a score closer to 1 means that less damage is detected	0.9992	0.0072
Underside_score	Car underside condition; a score closer to 1 means that less damage is detected	0.9991	0.0082

This study proposes the introduction of four new features: spare key, service book, warranty expired, and initial price. The initial price was selected for its capacity to characterize a vehicle and, theoretically, to exert an influence on the residual value of an asset. Three additional features, which also describe the condition of the vehicle, were incorporated into the model as they provide supplementary information that is relevant to the analysis [32]. In order to ascertain the factors influencing the decline in used car prices, this study has added a ratio of current and initial prices to the output variable. Subsequently, the performance of the models will be evaluated by means of a comparison of the RMSE, MAPE, and R^2 values. The application of SMBO to the optimization process yielded a combination of hyperparameters that demonstrated the potential to achieve optimal model performance. The four algorithms used as base learners were optimized. In the case of the RF algorithm, the max features hyperparameter determines the number of features that are taken into account in the construction of each tree, while the *n* estimators parameter defines the number of trees that are generated. The GB algorithm had hyperparameters analogous to those of the RF algorithm, with the additional feature of a learning rate, which regulated the contribution of each tree to the overall model performance. The SVR algorithm had three hyperparameters, namely C, epsilon, and gamma, which regulated the penalty size, tolerance margin, and kernel coefficient, respectively. Ultimately, the size of the NN algorithm was determined by the value of the hidden layer sizes parameter, which specifies the number of hidden layers and neurons for each hidden layer. Additionally, the alpha hyperparameter, which represents a regularization term, was subjected to tuning. The resulting combination of hyperparameters was then employed for modeling with a variety of machine learning algorithms based on the scikit-learn library in Python.

IV. RESULTS AND DISCUSSION

The findings indicated that the implementation of stacking and SMBO algorithms had a beneficial impact on the performance of both models. Therefore, these models demonstrate practical potential for accurately predicting used car prices and residual valuation ratios in real-world applications. The feature importance analysis revealed that the initial price of a vehicle has a significant impact on its resale value in a used condition. In addition, critical factors such as age, engine capacity, fuel type, brand, car type, mileage, and remaining warranty period have been identified as having a substantial impact on the determination of a vehicle's residual valuation ratio relative to its initial price.

A. Used Car Price Model

Table III presents the performance of the used car model, which demonstrates overall efficacy. Accordingly, the predictive abilities of all the models in this study were categorized as good after optimization, despite the fact that GB and SVR still exhibited considerable error before optimization. The models employing stacking algorithms demonstrated the most optimal performance among all the algorithms, as evidenced by their R^2 value of 0.9806, which was the highest. This conclusion was corroborated by the Mean Absolute Percentage Error (MAPE) value, which was 5.07% prior to model optimization, thereby indicating that the model exhibited an exceptionally high degree of predictive ability. Subsequently, the stacking algorithm was employed once more to model used car prices, with the hyperparameters of the base learner algorithm undergoing optimization. This resulted in an enhanced model performance, as evidenced by the elevated R^2

value and the reduced MAPE and Root Mean Square Error (RMSE). The R^2 value of 0.9836 indicated that the input variables exerted a direct influence on the variation in the output values within the model. Furthermore, the model exhibited a reduction in error, as indicated by a decline in the MAPE value to 4.84%. However, the MAPE value for the optimized model with the stacking algorithm was higher than that of the RF model. The consistently decreasing and lower RMSE values demonstrated that the stacking algorithm and optimization had a positive impact on the model's performance.

TABLE III. PERFORMANCE OF USED CAR PRICE MODEL

Metric	Algorithm	Without optimization	With SMBO optimization
RMSE	Stacking	14.99	13.79
	RF	19.18	14.25
	GB	31.35	14.75
	SVR	31.24	16.74
	NN	17.71	16.31
MAPE	Stacking	5.07%	4.84%
	RF	5.98%	4.65%
	GB	13.19%	5.17%
	SVR	10.63%	5.39%
	NN	5.88%	5.95%
R^2	Stacking	0.9806	0.9836
	RF	0.9683	0.9825
	GB	0.9153	0.9812
	SVR	0.9159	0.9758
	NN	0.9279	0.9771

As indicated in Table III, the model employing the stacking algorithm demonstrated superior performance compared to the other four algorithms across all indicators, including the R^2 value, both before and after SMBO optimization. This demonstrated that the model effectively mapped features for explaining the variations in the output values. Table III shows the impact of optimization on the performance of models constructed with single algorithms, such as SVR. Prior to optimization, the R^2 value of this model was comparatively low in comparison to the other models. However, following optimization, its performance became more closely aligned with that of ensemble models, including RF, GB, and stacking models. The positive effect of optimization on models comprising a combination of ML models, such as RF, GB, and stacking, was still evident, although to a lesser extent than previously observed. These findings corroborate the assertions of [33], who posited that ensemble learning could supplant hyperparameter optimization in an ML algorithm, and those of [34], who postulated that integrating hyperparameter optimization with ensemble learning could further enhance model performance.

Table III also demonstrates that MAPE and RMSE values were consistently lower for the stacking model, with the exception of the R^2 value, where the stacking model exhibited the lowest MAPE and RMSE values, outperforming the other models with the exception of the RF model, which had the lowest MAPE value after optimization. This indicated that the model with the stacking algorithm exhibited a lower degree of error when estimating used car prices. The impact of optimization on the MAPE and RMSE values corroborates prior findings on the enhancement of model performance through general optimization, as evidenced by the reduced MAPE and RMSE values for all algorithms post-optimization. Subsequent comparison revealed that the stacking model optimized through SMBO exhibited the most optimal performance for estimating used car prices, with R^2 = 0.9836, MAPE = 4.84%, and RMSE = 13.79. The results presented in Table III indicate that the proposed model, designed to predict the selling price of used cars, demonstrated superior performance compared to previous studies [6, 13]. The incorporation of initial price as a feature may have contributed to this outcome. In general, this feature can affect the residual value of an asset, thereby making it useful for predicting the selling price of used cars. Figure 2 confirms that the proposed model is highly reliant on this feature.

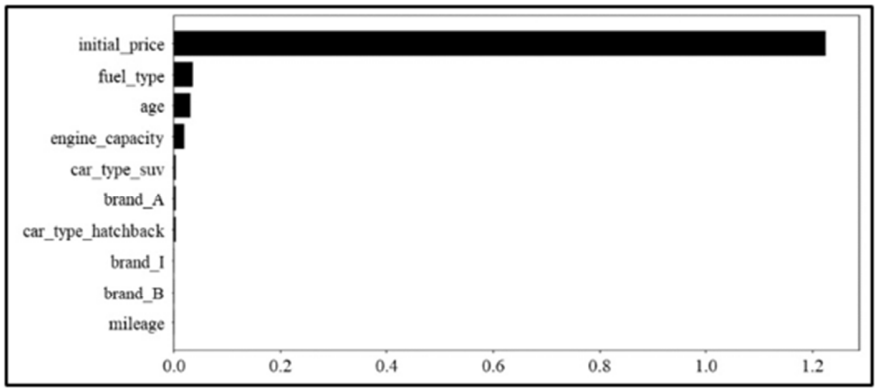

Fig. 2. Features with largest influence on used car price model.

B. Used Car Residual Valuation Ratio

The results of the used car residual valuation ratio model are presented in Table IV. The present model was evaluated under identical conditions as those used for the previous model. The performance of each of the five algorithms (RF, GB, SVR, NN, and stacking) was assessed with and without SMBO optimization. Prior to optimization, the RF-based used car residual valuation ratio model demonstrated superior performance with respect to the measurement indicators.

However, following optimization, the stacking model demonstrated the most optimal performance. As evidenced in Table IV, the optimized stacking model demonstrated superior performance compared to the other optimized models. This model exhibited the lowest RMSE and MAPE values, at 0.0400 and 3.98%, respectively, and the highest R^2 value, at 0.8061, while the optimized stacking model will serve as the reference point for assessing the significance of permutation features. The influence of the permutation feature was evaluated using the model to ascertain the most pivotal factors affecting the car valuation residual ratio. Figure 3 shows the features with the highest permutation importance score for the used car residual valuation ratio model constructed with the optimized stacking algorithm.

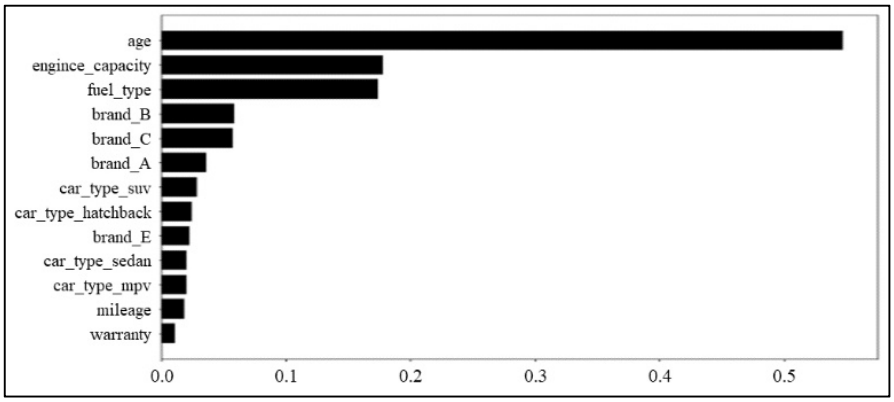

Fig. 3. Features with the greatest influence on used car residual valuation model.

As shown in Figure 3, the age of a used car (in months as of August 2022) had the greatest impact on the model's performance in terms of permutation importance. This finding provided compelling evidence of the dynamics of used car sales. In general, the value of a vehicle tends to decrease with age. Moreover, this finding was consistent with the results of previous studies [7, 12]. The following two features were identified as engine capacity and fuel type, which include the specifications of the vehicle being sold. It is therefore incumbent upon prospective buyers to be fully informed as to the specifications of the vehicle in question prior to making a purchase. It is therefore reasonable to conclude that these two features exert a significant influence on the selling price of a car. The following three features are associated with specific automotive brands: brand B, brand C, and brand A. These three Japanese automotive brands are highly regarded among Indonesian consumers. Japanese-manufactured cars are the most highly rated in Indonesia, due to their design aligning with Indonesian cultural preferences and offering superior resale value. Another feature meriting consideration is brand E. Despite not being as highly rated as the other three brands, brand E, a Japanese car brand, has consistently been among the top five brands (along with brand B, brand C, and brand A) in terms of sales in Indonesia over the past five years [4]. This suggests that brand B, brand C, brand A, and brand E are all significant variables to consider when estimating the selling price of a used car. The following two features describe the conditions of the used car: mileage and warranty expired. The feature "mileage" is defined as the distance a vehicle has traveled (in kilometers). It is often considered a factor that affects the selling price of a vehicle. In contrast, the "warranty expired" feature, which indicates the remaining warranty period of a vehicle in months, represents a novel discovery of this study. It is reasonable to assume that customers would desire to know how long a car company is financially responsible for any damages sustained by their vehicle. This information could prove valuable in future decision-making processes. The final four features are related to the specific type of vehicle in question. This finding is consistent with the conclusions of other research [13], which emphasized the influence of a vehicle's type on its market value.

TABLE IV. USED CAR RESIDUAL VALUATION RATIO MODEL PERFORMANCE

Metric	Algorithm	Without optimization	With SMBO optimization
RMSE	Stacking	0.0474	0.0400
	RF	0.0469	0.0416
	GB	0.0538	0.0412
	SVR	0.0594	0.0533
	NN	0.0547	0.0480
MAPE	Stacking	4.63%	3.98%
	RF	4.60%	4.04%
	GB	5.53%	4.14%
	SVR	5.75%	5.15%
	NN	5.42%	4.78%
R^2	Stacking	0.7275	0.8061
	RF	0.7339	0.7908
	GB	0.6488	0.7945
	SVR	0.5726	0.6565
	NN	0.6377	0.7207

V. CONCLUSIONS

In this study, two models were developed: a used car price model and a user car residual valuation ratio model. Both models were created by combining the stacking algorithm and Sequential Model-Based Optimization (SMBO). The models were trained using empirical data from the Indonesian used car market. The results of the Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2 measurements demonstrated that the models constructed with

the stacking algorithm and SMBO exhibited superior performance compared to other methods. A high-performing used car price model may be applied in the real world for the purpose of predicting used car prices. The feature importance analysis of this model corroborates the assertion that the initial price of a vehicle exerts a significant influence on its resale value. The second model was constructed with the objective of estimating the residual valuation ratio of a used car. The model also demonstrated satisfactory performance, thereby validating its suitability for estimating the factors influencing the ratio of the decline in used car prices. These factors can be classified into several categories. The initial category is the condition of the used car, which indicates the extent to which the previous owner used the vehicle. This includes the vehicle's age, mileage, and warranty status. Another category of factors pertains to the vehicle's main specifications, including engine capacity, fuel type, and body type. The final group is that of the brand, where Japanese-manufactured cars from brands such as Brand B, Brand C, Brand A, and Brand E exerted a significant influence on the ratio of price decline. While examining the models constructed with actual data using a combination of stacking and SMBO algorithms, it can be seen that the objectives of this study have been met.

The present study was limited to the examination of features pertaining to the characteristics of used cars, with the objective of estimating their valuation. In addition, the price of a used car may be influenced by external factors, including government regulations, developments in the automotive industry, and the time elapsed since the vehicle was sold. These factors were not examined in greater depth in the present study, but may be considered in future modelling processes. Furthermore, the number of samples can be augmented in future studies, as an increased sample size can enhance the learning process of a Machine Learning (ML) model.

ACKNOWLEDGMENT

This work was supported by Universitas Indonesia, which funded this study through the PUTI Q2 Research Grants Universitas Indonesia [grant number NKB-1337/UN2.RST/HKP.05.00/2022].

REFERENCES

[1] A. Jawed, M. A. H. Talpur, I. A. Chandio, and P. N. Mahesar, "Impacts of In-Accessible and Poor Public Transportation System on Urban Enviroment: Evidence from Hyderabad, Pakistan," *Engineering, Technology & Applied Science Research*, vol. 9, no. 2, pp. 3896–3899, Apr. 2019, https://doi.org/10.48084/etasr.2482.

[2] M. A. H. Talpur, M. Napiah, I. A. Chandio, T. A. Qureshi, and S. H. Khahro, "Development of a Regional Transport Policy Support System for Rural Planning Agencies in Developing World," *Procedia Engineering*, vol. 77, pp. 2–10, Jan. 2014, https://doi.org/10.1016/j.proeng.2014.07.003.

[3] *Number of Motor Vehicle by Type - Statistical Data*. Indonesia: BPS-Statistics, 2018.

[4] *Indonesian Automobile Industry Data*. Indonesia: GAIKINDO.

[5] D. J. *Bayu, Konsumen Lebih Pilih Beli Mobil Bekas Usai Pandemi*. Indonesia: Katadata, Nov. 19, 2020.

[6] F. R. Amik, A. Lanard, A. Ismat, and S. Momen, "Application of Machine Learning Techniques to Predict the Price of Pre-Owned Cars in Bangladesh," *Information*, vol. 12, no. 12, Dec. 2021, Art. no. 514, https://doi.org/10.3390/info12120514.

[7] E. Liu, J. Li, A. Zheng, H. Liu, and T. Jiang, "Research on the Prediction Model of the Used Car Price in View of the PSO-GRA-BP Neural Network," *Sustainability*, vol. 14, no. 15, Jan. 2022, Art. no. 8993, https://doi.org/10.3390/su14158993.

[8] C. Chen, L. Hao, and C. Xu, "Comparative analysis of used car price evaluation models," *AIP Conference Proceedings*, vol. 1839, no. 1, May 2017, Art. no. 020165, https://doi.org/10.1063/1.4982530.

[9] J.-D. Wu, C.-C. Hsu, and H.-C. Chen, "An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference," *Expert Systems with Applications*, vol. 36, no. 4, pp. 7809–7817, May 2009, https://doi.org/10.1016/j.eswa.2008.11.019.

[10] A. Wang, Q. Yu, X. Li, Z. Lu, X. Yu, and Z. Wang, "Research on Used Car Valuation Problem Based on Machine Learning," in *2022 International Conference on Computer Network, Electronic and Automation (ICCNEA)*, Xi'an, China, Sep. 2022, pp. 101–106, https://doi.org/10.1109/ICCNEA57056.2022.00032.

[11] S. Lessmann and S. Voß, "Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy," *International Journal of Forecasting*, vol. 33, no. 4, pp. 864–877, Oct. 2017, https://doi.org/10.1016/j.ijforecast.2017.04.003.

[12] E. Gegic, B. Isakovic, D. Keco, Z. Masetic, and J. Kevric, "Car Price Prediction using Machine Learning Techniques," *TEM Journal*, vol. 8, no. 1, pp. 113–118, Feb. 2019, https://doi.org/10.18421/TEM81-16.

[13] 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*, Cham, Switzerland: Springer, 2019, pp. 413–422, https://doi.org/10.1007/978-3-030-03402-3_28.

[14] P. Cerda and G. Varoquaux, "Encoding high-cardinality string categorical variables," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1164–1176, Mar. 2022, https://doi.org/10.1109/TKDE.2020.2992529.

[15] K. J. Liapis and D. D. Kantianis, "Depreciation Methods and Life-cycle Costing (LCC) Methodology," *Procedia Economics and Finance*, vol. 19, pp. 314–324, Jan. 2015, https://doi.org/10.1016/S2212-5671(15)00032-5.

[16] F. Pargent, F. Pfisterer, J. Thomas, and B. Bischl, "Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features," *Computational Statistics*, vol. 37, no. 5, pp. 2671–2692, Nov. 2022, https://doi.org/10.1007/s00180-022-01207-6.

[17] N. Zhang, Y. Su, B. Wu, X. Tu, Y. Jin, and X. Bao, "Cloud resource prediction model based on LSTM and RBF," in *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, Zhuhai, China, Sep. 2021, pp. 189–194, https://doi.org/10.1109/ICBASE53849.2021.00043.

[18] A. A. Alhashmi, A. M. Alashjaee, A. A. Darem, A. F. Alanazi, and R. Effghi, "An Ensemble-based Fraud Detection Model for Financial Transaction Cyber Threat Classification and Countermeasures," *Engineering, Technology & Applied Science Research*, vol. 13, no. 6, pp. 12433–12439, Dec. 2023, https://doi.org/10.48084/etasr.6401.

[19] M. Wistuba, N. Schilling, and L. Schmidt-Thieme, "Hyperparameter Search Space Pruning – A New Component for Sequential Model-Based Hyperparameter Optimization," in *Machine Learning and Knowledge Discovery in Databases*, Cham, Switzerland: Springer, 2015, pp. 104–119, https://doi.org/10.1007/978-3-319-23525-7_7.

[20] M. Massaoudi, S. S. Refaat, I. Chihi, M. Trabelsi, F. S. Oueslati, and H. Abu-Rub, "A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting," *Energy*, vol. 214, Jan. 2021, Art. no. 118874, https://doi.org/10.1016/j.energy.2020.118874.

[21] A. H. Alkenani, Y. Li, Y. Xu, and Q. Zhang, "Predicting Alzheimer's Disease from Spoken and Written Language Using Fusion-Based Stacked Generalization," *Journal of Biomedical Informatics*, vol. 118, Jun. 2021, Art. no. 103803, https://doi.org/10.1016/j.jbi.2021.103803.

[22] Y. A. Alsariera, M. H. Alanazi, Y. Said, and F. Allan, "An Investigation of AI-Based Ensemble Methods for the Detection of Phishing Attacks," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14266–14274, Jun. 2024, https://doi.org/10.48084/etasr.7267.

[23] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 1st ed. Boca Raton, FL, USA: Chapman & Hall/CRC, 2012.

[24] A. C. Faul, *A Concise Introduction to Machine Learning*, 1st ed. Boca Raton, FL, USA: Chapman and Hall/CRC, 2019.

[25] A. Callens, D. Morichon, S. Abadie, M. Delpey, and B. Liquet, "Using Random forest and Gradient boosting trees to improve wave forecast at a specific location," *Applied Ocean Research*, vol. 104, Nov. 2020, Art. no. 102339, https://doi.org/10.1016/j.apor.2020.102339.

[26] C. Antonio, "Sequential model based optimization of partially defined functions under unknown constraints," *Journal of Global Optimization*, vol. 79, no. 2, pp. 281–303, Feb. 2021, https://doi.org/10.1007/s10898-019-00860-4.

[27] "Jual Beli Mobil Bekas Terpercaya di Indonesia." Carsome, https://www.carsome.id/.

[28] A. A. Tanvir, I. A. Khandokar, A. K. M. Islam, S. Islam, and S. Shatabda, "A gradient boosting classifier for purchase intention prediction of online shoppers," *Heliyon*, vol. 9, no. 4, Apr. 2023, Art. no. e15163, https://doi.org/10.1016/j.heliyon.2023.e15163.

[29] J. Sill, G. Takács, L. Mackey, and D. Lin, "Feature-Weighted Linear Stacking," Nov. 2009.

[30] A. Hebbal, L. Brevault, M. Balesdent, E.-G. Talbi, and N. Melab, "Bayesian optimization using deep Gaussian processes with applications to aerospace system design," *Optimization and Engineering*, vol. 22, no. 1, pp. 321–361, Mar. 2021, https://doi.org/10.1007/s11081-020-09517-8.

[31] W. Gan, J. Li, and Y. Guo, "Research on ant colony optimization network access algorithm based on model of vehicle fog calculation, " in *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, Zhuhai, China, Sep. 2021, pp. 52–55, https://doi.org/10.1109/ICBASE53849.2021.00018.

[32] 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 (ICBIR)*, Bangkok, Thailand, May 2018, pp. 115–119, https://doi.org/10.1109/ICBIR.2018.8391177.

[33] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning: Methods, Systems, Challenges*, 1st ed.. Cham, Switzerland: Springer, 2019.

[34] J.-C. Lévesque, C. Gagné, and R. Sabourin, "Bayesian hyperparameter optimization for ensemble learning," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, Arlington, VA, USA, Jun. 2016, pp. 437–446.