

MDI and PI XGBoost regression-based methods: regional best pricing prediction for logistics services

Agus Purnomo¹, Aji Gautama Putrada², Roni Habibi³, Syafrianita⁴

¹Department of Master Logistics Management, Faculty of Logistics, Technology, and Business, Universitas Logistik dan Bisnis Internasional, Bandung, Indonesia

²Advanced and Creative, Networks Research Center, Telkom University, Bandung, Indonesia

³Department of Informatics Engineering, Faculty of Vocational Schools, Universitas Logistik dan Bisnis Internasional, Bandung, Indonesia

⁴Department of Transportation Management, Faculty of Logistics, Technology, and Business, Universitas Logistik dan Bisnis Internasional, Bandung, Indonesia

Article Info

Article history:

Received Feb 1, 2024

Revised May 15, 2024

Accepted May 26, 2024

Keywords:

Explainable artificial intelligence

Extreme gradient boosting

Mean decrease in impurity

Permutation importance

Retail price prediction

ABSTRACT

The logistics industry in Indonesia, with PT Pos Indonesia as the dominant player, is confronted with intense price competition. The challenge lies in establishing the most favorable price for regional logistics services in every region, with the aim of gaining a competitive edge and augmenting revenue. This intricate task encompasses local market conditions, competition, customer preferences, operational costs, and economic factors. To address this complexity, this study proposes the utilization of machine learning for price prediction. The price prediction model devised incorporates the extreme gradient boosting regression (XGBR), support vector machine (SVM), random forest, and logistics regression algorithms. This research contributes to the field by employing mean decrease in impurity (MDI) and permutation importance (PI) to elucidate how machine learning models facilitate optimal price predictions. The findings of this study can assist company management in enhancing their comprehension of how to make informed pricing decisions. The test results demonstrate values of 0.001, 0.005, 0.458, 0.009, and 0.9998. By employing machine learning techniques and explanatory models, PT Pos Indonesia can more accurately determine optimal prices in each region, bolster profits, and effectively compete in the expanding regional market.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Agus Purnomo

Department of Master Logistics Management, Faculty of Logistics, Technology, and Business

Universitas Logistik Dan Bisnis Internasional

St. Sari Asih No.54, Bandung, West Java, Indonesia

Email: aguspurnomo@ulbi.ac.id

1. INTRODUCTION

The influence of digital technology developments is a challenge for the largest logistics company in Indonesia, namely PT Pos Indonesia, the challenges faced by PT Pos Indonesia can compete globally to provide the best service, one of which is determining the regional best pricing, referring to determining the best price for products or services in various regions or different geographical areas [1]. The main objective of determining the regional best pricing is to optimize the company's revenue and profits by considering relevant factors such as local market conditions, competition, customer demand, operational costs, and other relevant factors [2].

Digital technology has changed how customers interact with brands and products, influencing their preferences and purchasing decisions. Therefore, understanding local consumer behavior is key to determining

optimal prices in various regions. Many recent research have discussed similar problems. Rickert *et al.* [3] mentioned that determining regional best pricing involves analyzing the data and information available for each region, including sales data, competitor prices, customer demographics, purchasing power levels, local preferences, and other economic factors. Phillips [4] said that determining regional best pricing often involves a combination of global price standards applied by the company and price adjustments specific to each region. Palmatier and Sridhar [5] used factors such as cost differences, level of competition, customer preferences, and local policies, which influenced price adjustments made in each region. However, Zhang [6] stated that this isn't easy to do based on the complex and fluctuating data characteristics, so computing is needed so that every need in determining the optimal best price is based on relevant factors.

Based on previous research, Chen *et al.* [7] used extreme gradient boosting (XGBoost), a machine learning approach to create optimal price prediction modeling in each region based on relevant factors. According to Zheng *et al.* [8], machine learning can help identify important pricing factors and provide appropriate price recommendations for each region using a regression approach. The features used in the research of Akyildirim *et al.* [9] include demographic variables, competitor prices, geographic data, customer preferences, and other factors related to pricing. Ullah *et al.* [10] used the mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) metrics to measure the performance of their prediction model. Next, Jain *et al.* [11] used the best machine learning models to make optimal price predictions for each region based on relevant factors. Machine learning can solve complex problems such as local consumer behavior analysis. However, the problem is that the results are difficult for humans to interpret [10]. Fumagalli *et al.* [12] showed that permutation importance (PI) is another useful method for explainable artificial intelligence (XAI).

The best pricing of PT Pos Indonesia's logistics services in different regions could be more optimal, so it loses sales competition with other logistics providers. The non-optimal price is because PT Pos Indonesia has not used the best price prediction method in determining the regional best prices, which includes relevant factors such as the variable total price of competitors, the number of customers, the freight price, the number of competitors and the product score given by the customer. As a result, the price of logistics services set from period to period becomes uncompetitive and loses competition with the prices of other logistics service providers. Therefore, the problem in this study is how to create the best logistics service price prediction model by including relevant factors that can be used for each region of PT Pos Indonesia to compete with other logistics service providers. Inspired by the gap that has been explained, our research aims to create the best regional price prediction for logistics services by including relevant factors with a high level of explanation so that PT Pos Indonesia can be competitive with other logistics providers.

We suggest using mean decrease in impurity (MDI) and PI to provide the expected level of explanation based on relevant factors. We used XGBoost regression (XGBR) to predict the price of logistics services and compared it with eight other regression models. The model performance is then evaluated using R-squared, MSE, RMSE, or MAPE. Furthermore, the best machine learning model is used to make optimal price predictions for each region based on relevant factors. This research uses local consumer behavior data analysis and machine learning approaches to help companies such as PT Pos Indonesia understand consumer preferences and behavior in different regions. Finally, we use MDI and PI methods to improve the interpretation of PT Pos Indonesia's local consumer behavior analysis. To the best of our knowledge, no research uses machine learning to analyze local consumer behavior for optimal regional pricing, especially for logistics service providers. The contributions of this research are, therefore, as follows:

- a. To create an optimal price prediction model for logistics services using XGBR that can be applied in different regions so that the company can be competitive with its competitors in terms of price.
- b. Produce a logistics services price forecasting model that can be explained by MDI and PI, illustrating the sensitivity of the model to various relevant factors.

The remainder of this paper is presented in several parts. Section 2 contains a review of previous research that supports our findings and the research design and method. The results of our research are presented in section 3. Finally, section 4 summarises and highlights the main points of our contribution.

2. METHOD

This research aims to determine the optimal regional price using a machine learning model. Machine learning involves creating and adapting models for data analysis, which allows programs to learn through experience. Jain *et al.* [11] performed price prediction using regression models and support vector machine (SVM), with a high degree of accuracy. In addition, machine learning approaches are also effective for price comparison. Derdouri and Murayama [13] conducted a comparative study of price estimation using the machine learning random forest model, achieving 79% accuracy, with RMSE 0.1537 and MAE 0.1139. Gu *et al.* [14] achieved 99.1% accuracy with the random forest regression (RFR) model, while Mohd *et al.* [15] achieved 92.4%

for SVM. In addition to machine learning methods, price prediction can also be done using statistical approaches. Lu *et al.* [16] use a stepwise regression statistical model for price prediction.

We propose a research method consisting of several steps, as shown in Figure 1. First, we collect the best pricing data. The next step is the data pre-processing stage to identify missing values and transformation. After data pre-processing is complete, regression modeling will be carried out to determine the best pricing based on the data and each specific factor throughout the region. Then, the results of the modeling are evaluated for performance. The final step is to report the research results.

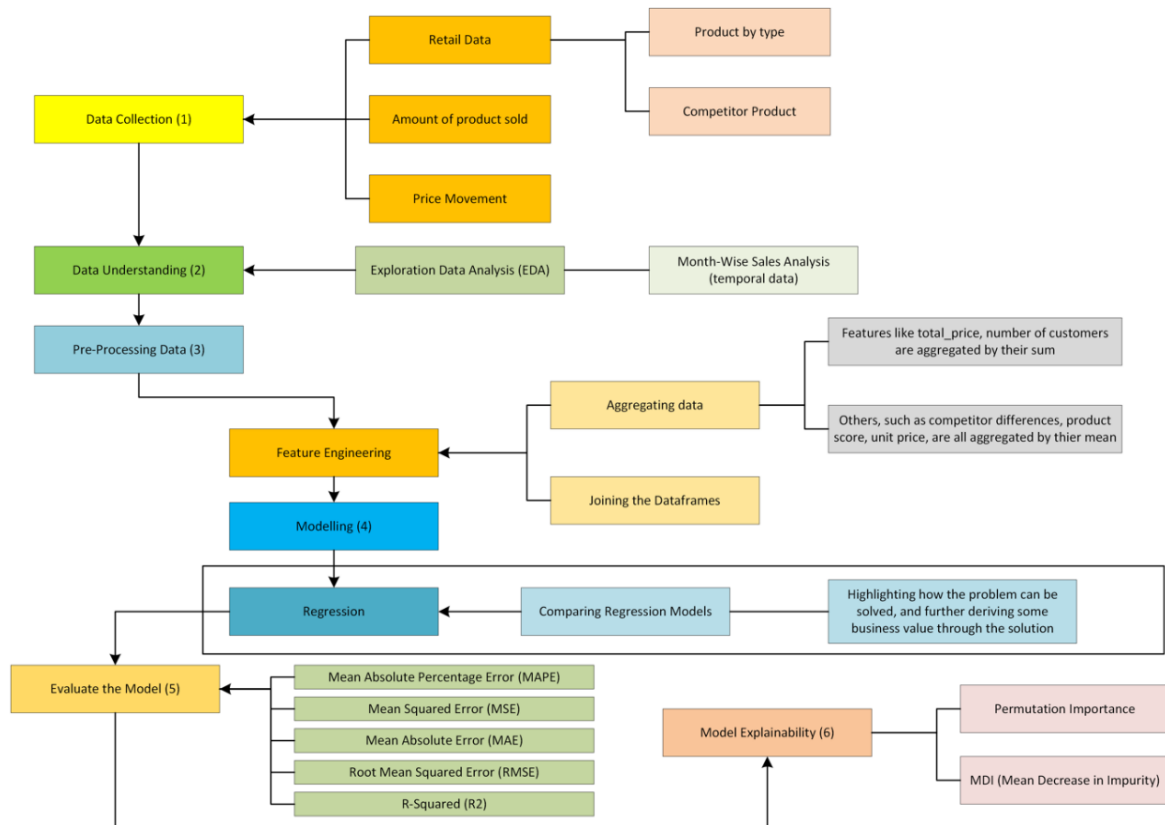


Figure 1. Proposed method

2.1. The retail price dataset

Data collection is collecting information or data from various sources for analysis, research, decision-making, or other purposes [17]. This process is essential in scientific research, providing the foundation for testing hypotheses and drawing conclusions. In business analysis, data collection helps organizations understand market trends and customer behaviour. Similarly, it is crucial in product development, project management, and numerous other fields where informed, evidence-based decisions are necessary.

2.2. Data understanding

Furthermore, data understanding is one of the important stages in the data analysis process. This involves exploration and initial understanding of the data used in the analysis. The goal of this stage is to gain a better understanding of the data, including its characteristics, structure, and context.

2.3. Pre-processing

The data pre-processing stage is one of the most important stages in data analysis. This is the process of preparing data before the data can be used for analysis or modeling. The goal is to clean, tidy, and organize data so that the data becomes more useful and ready to be used in statistical analysis or modeling [18]. Using the average to fill in missing values by replacing the missing values with the average of all data in the dataset column used. The formula is as (1):

$$\mu_V = \frac{\sum_{n \in N} V_n}{N} \quad (1)$$

where μ_V is the final result of the data sought, namely the average value of a data group. Then $\sum_{n \in N} V_n$ is the total of all values in the data group. Add all the numbers together to get this value. Finally, N is the total number of values in the data group, measuring how much data one has.

2.4. The prediction model

At the modeling stage, we carry out retail price prediction using XGBR. We benchmark it with various regression models such as linear regression (LR), ridge regression (RR), Lasso, RFR, gradient boosting regression (GBR), AdaBoost regression (ABR), K-nearest neighbor regression (KNN), and super vector regression (SVR). Regression is a statistical technique to analyze the relationship between two or more variables. LR is a technique in statistics and machine learning that is used to model linear relationships between one or more independent variables (predictors) and dependent variables (targets) [19]. RR is a linear regression technique used in statistics and regression analysis. This is a method used to overcome multicollinearity problems in regression analysis [20], where the independent variables in the regression model have a high correlation with each other. This ridge penalty term can also be used for feature selection in some other research. Lasso regression is a linear regression method used to overcome multicollinearity problems (when two or more independent variables are highly correlated) and influence feature selection in the model [21]. RFR is a method in machine learning that is used to carry out regression or predictions [22]. GBR is an ensemble algorithm in machine learning used for regression modeling [23]. ABR is an ensemble algorithm in machine learning used for regression modeling [24]. KNN is a regression algorithm used to predict the value of a dependent variable based on the independent variable values of K nearest neighbors in the [25] training dataset. This is a simple but effective method in case of regression. SVR is a regression algorithm that uses the concept of SVM to make predictions on regression data [26]. XGBR is a type of ensemble learning, namely a machine learning method that uses several models simultaneously to improve [27] performance. XGBR specifically is a boosting type learning ensemble, namely ensemble learning that lines up weak learners serially, where each weak learner reduces the error from the previous weak learner by optimizing the loss function with the Newton-Raphson method [28].

2.5. Feature engineering

Feature engineering is a process in which a data scientist or data engineer creates new features (variables) or changes existing features in a dataset to improve the quality of a machine learning model [29]. The main goal of feature engineering is to create more informative datasets, reduce noise, and enable models to understand patterns in the data better. Feature engineering is very important in machine learning because good features can have a big impact on model quality.

2.6. Evaluation metric

The approach used in the final step of this research is based on general statistical assessments, namely MSE, RMSE, and MAPE [30]. The MSE, RMSE, MAE, MAPE, and R-squared assessments measure the model's effectiveness in predicting best pricing. MSE, RMSE, MAE, and MAPE assess model performance, while R-squared assesses the model's predictive ability.

2.7. The explainability model

The goal of model explainability is to explain how a machine-learning model makes decisions so that the model being used can be understood and trusted [31]. In this research, two methods are MDI and PI, to achieve this goal. MDI is a measure commonly used in decision tree-based models, such as random forests. It assesses the importance of a feature by evaluating how much the feature contributes to reducing the impurity or uncertainty in the model's predictions. In other words, MDI helps identify which features are most influential in making accurate predictions. At each node in a decision tree, we measure the Gini impurity, which is a metric indicating how mixed the target classes are within that node. The formula for Gini impurity at a node t with K classes is:

$$Gini(t) = 1 - \sum_{i=1}^K (p(i|t))^2 \quad (2)$$

where $p(i|t)$ represents the proportion of samples from class i within node t . For a specific feature, MDI is computed by averaging the reduction in Gini impurity across all nodes that use that feature to make decisions. For instance, if feature X is used in n nodes and $Gini(t)$ is the Gini impurity at node t , then the MDI for feature X is expressed as (3):

$$MDI(X) = \frac{1}{n} \sum_t Gini(t) \quad (3)$$

A higher MDI indicates that the feature is more crucial in making predictions within the random forest model. PI is a method used to measure how important each feature is in a machine-learning model [32]. It observes a feature by randomly permuting the feature's values. If the model performance fluctuates, the model is sensitive towards that feature and insensitive if otherwise [33]. The main goal is to provide an understanding of the relative contribution of each feature to the model's ability to make predictions. Let the score baseline be the model performance score on the original data (without permutation), and the score permuted be the model performance score after permuting a particular feature. If N is the number of permutations conducted, then the PI for a feature X can be computed using the formula:

$$PI(X) = \frac{1}{N} \sum_{i=1}^N (score_{permuted}^{(i)} - score_{baseline}) \quad (4)$$

where i is the permutation index.

3. RESULTS AND DISCUSSION

3.1. Result

In the first test, we analyzed the retail price periodically. Here, the total price is aggregated monthly by summing every value. We plot a regression line between total price and aggregate customers (also summed up monthly) from the dataset. This can model the relationship between these two variables. Figure 2 shows the linear relationship between total price and customers. We can interpret the linear relationship objectively with the R-squared value, which is 0.98. That number is considered very high because it approximates the best value of R-squared, 1.0. The p-value of the regression line is 0.01, meaning the null hypothesis is rejected. In regression analysis, rejecting the null hypothesis means there is a significance in the slope of the regression line and that the two variables are strongly related. In normative terms, the increase in total price is related to the increase of customers that visit the store.



Figure 2. Linear regression and customer per-month analysis; total price vs customers

In the second test, we plot a regression line between the total price and the number of weekends per month from the dataset to observe the relationship between these two variables. Figure 3 shows the linear relationship between total price and customers. The R-squared value of the regression line is 0.86. That number is considered high, however, it is not as high as the previous result. On the other hand, the p-value of the regression line is 0.01, meaning the null hypothesis is also rejected, which leads to the conclusion that the null hypothesis is rejected. There is still a significance in the slope of the regression line, while the two variables are strongly related. In normative terms, the increase in total price is correlated to the increase in the number of weekends per month of retail.

Analysis of the customer per-month bar chart can be useful for several things, including trend analysis, churn, market effectiveness, customer planning, and prediction. Figure 4 shows that the most customers were in November 2017, and the fewest were in January 2017. After conducting data exploration, we carry out the data pre-processing stage. At this stage, we group data between average and total. The data grouped to calculate the average is 'product id,' 'month year,' 'comp1 diff,' 'comp2 diff,' 'comp3 diff,' 'fp1 diff,' 'fp2 diff,' 'fp3 diff,' 'product score,' and 'unit price.' Meanwhile, the data that is grouped to calculate the total amount is 'product id,' 'month year,' 'total price,' 'freight price,' and 'customers'.

After the data has been grouped into average and total, the next step is to calculate the average and total of the two data groups. The results of these calculations are stored in two variables, namely, product mean and product sum. Next, after getting the average and total results, the two are combined into one data frame,

which contains information about the average and total based on ‘product id’. The final stage in the feature engineering process is calculating the logarithm of the variable to be predicted, namely ‘unit price,’ and the results will be stored in the variable y log, which contains the logarithm values from ‘unit price’. The next step is modelling. At this stage, eight regression models are compared to get the best prediction value. The Table 1 displays the evaluation value of each regression model. The XGBR model has the best results compared to other regression models, with an MSE value of 0.0001, MAE of 0.005, MAPE of 0.458, RMSE of 0.009, and R-square of 0.9998.



Figure 3. Linear regression and customer per-month analysis; Weekly analysis of total



Figure 4. Linear regression and customer per-month analysis; customer per-month

Table 1. Regression model performance comparison

Model	Evaluation metrics				
	MSE	MAE	MAPE	RMSE	R2
LR	0.1121	0.258	28.193	0.335	0.7243
RR	0.1127	0.263	28.082	0.336	0.7229
Lasso	0.1525	0.333	35.971	0.390	0.6250
RFR	0.0149	0.101	10.212	0.122	0.9633
GBR	0.0016	0.031	3.117	0.040	0.9961
ABR	0.0148	0.097	9.728	0.122	0.9645
XGBR	0.0001	0.005	0.458	0.009	0.9998
KNR	0.1008	0.253	25.286	0.318	0.7520
SVR	0.1503	0.294	34.944	0.388	0.6302

The explainability model stage is to explain and describe why the XGBR model produces certain decisions and results. The MDI graph in Figure 5 shows the relationship between feature values and their impact on predicted values. MDI values on the y-axis (vertical) play a crucial role in understanding the significance of features in the prediction-making process. The elevation of MDI values indicates the level of importance each feature holds in influencing predictions. Visualized as bars on the graph, each feature’s bar height signifies the magnitude of its impact. The emphasis should be placed on bars with the highest MDI values, as these features are deemed the most pivotal in shaping the model’s predictions. Features characterized by elevated MDI values play a substantial role in minimizing impurity during the construction of decision trees. Furthermore, a positive MDI value signifies a positive correlation between the feature and the prediction outcome. In simpler terms, higher values of the feature generally support higher predictions. This analysis aids in comprehending the pivotal features that contribute significantly to the accuracy of the model’s predictions.

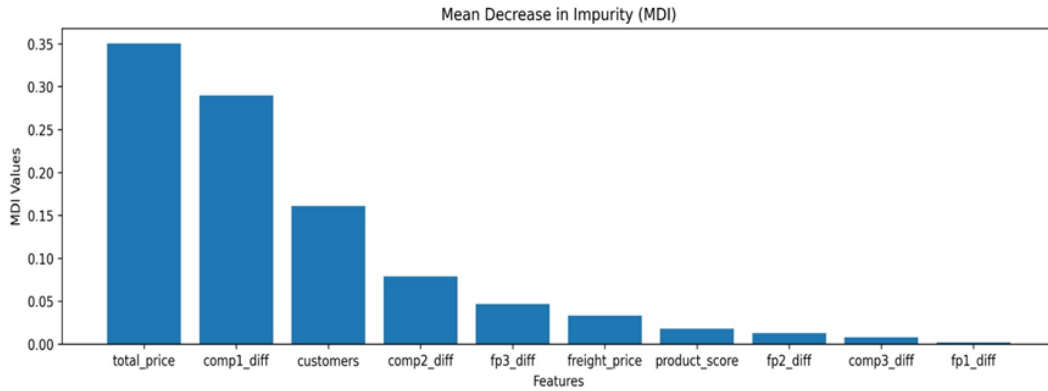


Figure 5. MDI summary plot

The PI value serves as a valuable metric for understanding the impact of features on a model’s performance when their values are randomly permuted. In Figure 6 and Table 2, we present the PI results, wherein feature names are arranged based on their weight magnitudes. Subsequently, we compute the maximum and minimum error values derived from the PI analysis. It is noteworthy that the features of the highest PI weight correspond with the MDI, specifically the ‘total price’. This consistent alignment indicates that ‘total price’ significantly influences both MDI and PI, underscoring its importance in predicting outcomes.

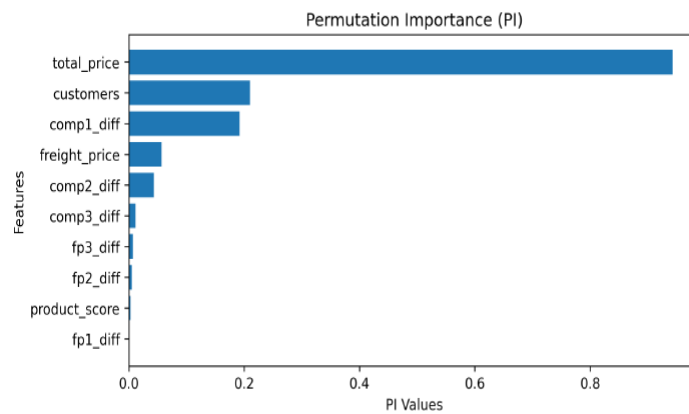


Figure 6. PI summary plot

However, disparities arise when comparing the lowest weight in PI with the MDI score. While PI identifies ‘fp1 diff’ as the feature with the lowest weight, MDI designates ‘freight price’ as having the lowest MDI value. This discrepancy highlights the nuanced nature of PI, which is inherently model-specific. In this particular instance, the model under examination is XGBoost (XGBR), revealing that PI results can be influenced by the intricacies of the underlying model. A comprehensive understanding of these differences enhances our insight into how features contribute to model performance, taking into account both MDI and PI perspectives.

Table 2. The PI result

Weight	Feature
0.9896±0.2898	‘total’
0.1006±0.0450	‘comp2’
0.0966±0.0172	‘comp1’
0.0845±0.0341	‘customers’
0.0302±0.0231	‘comp3’
0.0123±0.0025	‘product’
0.0119±0.0057	‘fp2’
0.0084±0.0013	‘fp3’
0.0048±0.0022	‘freight’

3.2. Discussion

Several studies have carried out price predictions with various regression models. Shahrel *et al.* [34] proved that SVR is better than linear regression in price prediction. Durganjali and Pujitha [35] proposed ABR for house price prediction. Finally, Bonamutial and Prasetyo [36] demonstrated that RFR is better than KNR in smartphone price prediction. Our research shows that XGBR has better overall performance than LR, RR, Lasso, RFR, ABR, KNR, and SVR in predicting retail prices. Our research contribution is an optimum retail price prediction using XGBR.

In our research, we highlight the different interpretations offered by MDI and PI techniques in analyzing features. These interpretations, when combined, contribute to a more comprehensive XAI. While both methods score features based on their contribution to predictive power, MDI goes a step further by providing insight into the positive/negative impact of each feature. In contrast, PI offers error values that indicate the sensitivity of features and their influence on overfitting. Our dual contribution is to improve the explanation of retail price prediction models through PI techniques and to leverage MDI and PI to analyze influential features, especially ‘total price’. Our analysis underscores the important role of ‘total price,’ ‘comp1 diff,’ and ‘customer’ in model development, with ‘total price’ being the most influential. The choice between MDI and PI depends on the research needs. Overall, the findings emphasize the importance of ‘total price’ in forming an optimized pricing model. Our research contributions are summarised in Table 3 by comparing them with state-of-the-art research in retail price prediction.

Table 3. A comparison of state-of-the-art research on the retail price prediction

Reference	Prediction model	R-squared	Explainability method	
			MDI	PI
Shahrel <i>et al.</i> [34]	SVR	0.6302	✘	✘
Durganjali and Pujitha [35]	ABR	0.9645	✘	✘
Bonamutial and Prasetyo [36]	RFR	0.9633	✘	✘
Proposed method	XGBR	0.9998	✓	✓

4. CONCLUSION

In this study, a regional best price prediction model for logistics services was successfully constructed using the XGBR and its accompanying explanation model. The aim was to enhance the interpretability of the price prediction. To compare the effects of various factors on the model’s analysis, XGBR was tested against LR, RR, Lasso, RFR, GBR, ABR, KNR, and SVR. Additionally, two XAI methods, namely MDI and PI, were employed. The results of the tests revealed that XGBR surpassed the benchmark method. This was corroborated by the MSE, MAE, MAPE, RMSE, and R-squared values, which were found to be 0.0001, 0.005, 0.458, 0.009, and 0.9998, respectively. Furthermore, based on the MDI and PI explanatory models, it was determined that total price was the most influential factor in predicting the optimal regional best price for logistics services.

This study demonstrates the superiority of XGBR over eight other regression models, establishing it as the most effective approach. Moreover, it holds practical implications for the logistics industry, enabling companies to determine the optimal regional best pricing prediction for logistics services. This knowledge gives them a competitive advantage over their rivals, leading to increased sales and profits.

This study proposes two potential directions for further academic inquiry aimed at enhancing the precision of regional best pricing prediction as a strategy for gaining a competitive advantage in the logistics industry. Firstly, we advocate for the utilization of datasets sourced directly from the company in order to enhance the accuracy and applicability of the model to the company’s specific circumstances. In terms of model development, we suggest integrating constraint functions to accommodate intricate decision-making strategies in order to determine the optimal regional best pricing prediction for logistics services. This objective can be accomplished by implementing evolutionary algorithms, which enable the model to adapt to fluctuations in market dynamics.

ACKNOWLEDGEMENTS

All authors would like to thank Universitas Logistik Dan Bisnis Internasional (ULBI). Then also to Mr. Harsh Singh who has published the data on Kaggle. then to the supporting colleagues in this research namely Mr. Rofi Nafiis Zain and Mr. Bahtiar Ramadhan as applied undergraduate study program students of informatics engineering.




REFERENCES

- [1] D. L. Huff, "Defining and estimating a trading area," *Journal of Marketing*, vol. 28, no. 3, pp. 34–38, 1964, doi: 10.1177/002224296402800307.
- [2] J. Lintner, "Dividends, earnings, leverage, stock prices and the supply of capital to corporations," *The Review of Economics and Statistics*, vol. 44, no. 3, pp. 243–269, 1962, doi: 10.2307/1926397.
- [3] D. Rickert, J. P. Schain, and J. Stiebale, "Local market structure and consumer prices: Evidence from a retail merger," *The Journal of Industrial Economics*, vol. 69, no. 3, pp. 692–729, 2021, doi: 10.1111/joie.12275.
- [4] R. L. Phillips, *Pricing and revenue optimization*. Stanford university press, 2021.
- [5] R. W. Palmatier and S. Sridhar, *Marketing strategy: Based on first principles and data analytics*. Bloomsbury Publishing, 2020.
- [6] X. Zhang, C. Zhang, and Z. Wei, "Carbon Price Forecasting Based on Multi-Resolution Singular Value Decomposition and Extreme Learning Machine Optimized by the Moth-Flame Optimization Algorithm Considering Energy and Economic Factors," *Energies*, vol. 12, no. 22, p. 4283, Nov. 2019, doi: 10.3390/en12224283.
- [7] L. Chen *et al.*, "Measuring impacts of urban environmental elements on housing prices based on multisource data—a case study of Shanghai, China," *ISPRS International Journal of Geo-Information*, vol. 9, no. 2, p. 106, 2020, doi: 10.3390/ijgi9020106.
- [8] Q. Zheng, J. Chen, R. Zhang, and H. H. Wang, "What factors affect Chinese consumers' online grocery shopping? Product attributes, e-vendor characteristics and consumer perceptions," *China Agricultural Economic Review*, vol. 12, no. 2, pp. 193–213, 2020, doi: 10.1108/CAER-09-2018-0201.
- [9] E. Akyildirim, A. Goncu, and A. Sensoy, "Prediction of cryptocurrency returns using machine learning," *Annals of Operations Research*, vol. 297, pp. 3–36, 2021, doi: 10.1007/s10479-020-03575-y.
- [10] I. Ullah, K. Liu, T. Yamamoto, M. Zahid, and A. Jamal, "Modeling of machine learning with SHAP approach for electric vehicle charging station choice behavior prediction," *Travel Behaviour and Society*, vol. 31, pp. 78–92, 2023, doi: 10.1016/j.tbs.2022.11.006.
- [11] M. Jain, H. Rajput, N. Garg, and P. Chawla, "Prediction of House Pricing using Machine Learning with Python," *2020 International Conf. on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 570–574, doi: 10.1109/ICESC48915.2020.9155839.
- [12] F. Fumagalli, M. Muschalik, E. Hüllermeier, and B. Hammer, "Incremental permutation feature importance (iPFI): towards online explanations on data streams," *Machine Learning*, vol. 112, no. 12, pp. 4863–4903, 2023, doi: 10.1007/s10994-023-06385-y.
- [13] A. Derdouri and Y. Murayama, "A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across Fukushima prefecture, Japan," *Journal of Geographical Sciences*, vol. 30, pp. 794–822, 2020, doi: 10.1007/s11442-020-1756-1.
- [14] S. Gu, B. Kelly, and D. Xiu, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, vol. 33, no. 5, pp. 2223–2273, May 2020, doi: 10.1093/rfs/hhaa009.
- [15] T. Mohd, S. Masrom, and N. Johari, "Machine learning housing price prediction in Petaling Jaya, Selangor, Malaysia," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2, pp. 542–546, 2019.
- [16] H. Lu, X. Ma, K. Huang, and M. Azimi, "Carbon trading volume and price forecasting in China using multiple machine learning models," *Journal of Cleaner Production*, vol. 249, p. 119386, 2020, doi: 10.1016/j.jclepro.2019.119386.
- [17] L. E. Tomaszewski, J. Zaretsky, and E. Gonzalez, "Planning qualitative research: Design and decision making for new researchers," *International Journal of Qualitative Methods*, vol. 19, 2020, doi: 10.1177/1609406920967174.
- [18] M. K. Shende, A. E. Feijoo-Lorenzo, and N. D. Bokde, "cleanTS: Automated (AutoML) tool to clean univariate time series at microscales," *Neurocomputing*, vol. 500, pp. 155–176, 2022, doi: 10.1016/j.neucom.2022.05.057.
- [19] F. Fang *et al.*, "Cryptocurrency trading: a comprehensive survey," *Financial Innovation*, vol. 8, no. 1, pp. 1–59, 2022, doi: 10.1186/s40854-021-00321-6.
- [20] S. Çankaya, S. Eker, and S. H. Abac, "Comparison of Least Squares, Ridge Regression and Principal Component approaches in the presence of multicollinearity in regression analysis," *Turkish Journal of Agriculture-Food Science and Technology*, vol. 7, no. 8, pp. 1166–1172, 2019, doi: 10.24925/turjaf.v7i8.1166-1172.2515.
- [21] B. Liu, Y. Jin, D. Xu, Y. Wang, and C. Li, "A data calibration method for micro air quality detectors based on a LASSO regression and NARX neural network combined model," *Scientific Reports*, vol. 11, no. 1, p. 21173, 2021, doi: 10.1038/s41598-021-00804-7.
- [22] E. M. M. der Heide, R. F. Veerkamp, M. L. V. Pelt, C. Kamphuis, I. Athanasiadis, and B. J. Ducro, "Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle," *Journal of Dairy Science*, vol. 102, no. 10, pp. 9409–9421, Oct. 2019, doi: 10.3168/jds.2019-16295.
- [23] S. F. Pane, A. G. Putrada, N. Alamsyah, and M. N. Fauzan, "A PSO-GBR Solution for Association Rule Optimization on Supermarket Sales," *2022 Seventh International Conference on Informatics and Computing (ICIC)*, Denpasar, Bali, Indonesia, 2022, pp. 1–6, doi: 10.1109/ICIC56845.2022.10007001.
- [24] B. Liu, C. Liu, Y. Xiao, L. Liu, W. Li, and X. Chen, "AdaBoost-based transfer learning method for positive and unlabelled learning problem," *Knowledge-Based Systems*, vol. 241, 2022, doi: 10.1016/j.knsys.2022.108162.
- [25] N. A. Sami and D. S. Ibrahim, "Forecasting multiphase flowing bottom-hole pressure of vertical oil wells using three machine learning techniques," *Petroleum Research*, vol. 6, no. 4, pp. 417–422, 2021, doi: 10.1016/j.ptlrs.2021.05.004.
- [26] M. Zulfiqar, M. Kamran, M. B. Rasheed, T. Alquthami, and A. H. Milyani, "Hyperparameter optimization of support vector machine using adaptive differential evolution for electricity load forecasting," *Energy Reports*, vol. 8, pp. 13333–13352, 2022, doi: 10.1016/j.egy.2022.09.188.
- [27] M. Abdurrohmam and A. G. Putrada, "Forecasting Model for Lighting Electricity Load with a Limited Dataset using XGBoost," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 8, no. 2, 2023, doi: 10.22219/kinetik.v8i2.1687.
- [28] A. G. Putrada, N. Alamsyah, S. F. Pane, and M. N. Fauzan, "XGBoost for IDS on WSN Cyber Attacks with Imbalanced Data," *2022 International Symposium on Electronics and Smart Devices (ISESD)*, Bandung, Indonesia, 2022, pp. 1–7, doi: 10.1109/ISESD56103.2022.9980630.
- [29] A. Popov, "Feature engineering methods," in *Advanced Methods in Biomedical Signal Processing and Analysis*, 2023, pp. 1–29, doi: 10.1016/B978-0-323-85955-4.00004-1.
- [30] S. Kumar, T. Kolekar, K. Kotecha, S. Patil, and A. Bongale, "Performance evaluation for tool wear prediction based on Bi-directional, Encoder-Decoder and Hybrid Long Short-Term Memory models," *International Journal of Quality & Reliability Management*, vol. 39, no. 7, pp. 1551–1576, 2022, doi: 10.1108/IJQR-08-2021-0291.
- [31] D. Lundstrom and M. Razaviyayn, "Distributing Synergy Functions: Unifying Game-Theoretic Interaction Methods for Machine-Learning Explainability," *arXiv*, 2023, doi: 10.48550/arXiv.2305.03100.




- [32] C. Molnar, T. Freiesleben, G. König, G. Casalicchio, M. N. Wright, and B. Bischl, "Relating the partial dependence plot and permutation feature importance to the data generating process," *World Conference on Explainable Artificial Intelligence*, 2023, pp. 456-479, doi: 10.1007/978-3-031-44064-9_24.
- [33] A. G. Putrada, M. Abdurohman, D. Perdana, and H. H. Nuha, "EdgeSL: Edge-Computing Architecture on Smart Lighting Control with Distilled KNN for Optimum Processing Time," *IEEE Access*, vol. 11, pp. 64697-64712, 2023, doi: 10.1109/ACCESS.2023.3288425.
- [34] M. Z. Shahrel, S. Mutalib, and S. Abdul-Rahman, "PriceCop-Price Monitor and Prediction Using Linear Regression and LSVM-ABC Methods for E-commerce Platform," *International Journal of Information Engineering & Electronic Business*, vol. 13, no. 1, 2021, doi: 10.5815/ijieeb.2021.01.01.
- [35] P. Durganjali and M. V. Pujitha, "House Resale Price Prediction Using Classification Algorithms," *2019 International Conference on Smart Structures and Systems (ICSSS)*, Chennai, India, 2019, pp. 1-4, doi: 10.1109/ICSSS.2019.8882842.
- [36] M. Bonamutial and S. Y. Prasetyo, "Exploring the Impact of Feature Data Normalization and Standardization on Regression Models for Smartphone Price Prediction," *2023 International Conference on Information Management and Technology (ICIMTech)*, Malang, Indonesia, 2023, pp. 294-298, doi: 10.1109/ICIMTech59029.2023.10277860.

BIOGRAPHIES OF AUTHORS






Agus Purnomo    is a full time Associate Professor at the Department of Master Logistics Management, Universitas Logistik Dan Bisnis Internasional. His research experience is in the field of logistics and supply chain management. He earned his Ph.D. degree in 2009 at the Universitas Padjadjaran Bandung in the field of Operations Management. Master's degree in 1997 at the Institut Teknologi Bandung majoring in Industrial Engineering. Bachelor's degree in 1989 at Universitas Pasundan Bandung majoring in Industrial Engineering. He can be contacted at email: aguspurnomo@ulbi.ac.id.






Aji Gautama Putrada    received a Bachelor's degree in Electrical Engineering ITB 2008 and a Master's degree in Microelectronics ITB 2013. He became a lecturer at Telkom University, Bandung, and is currently an Assistant Professor. Since 2015, he has been involved in various research grants from the Government about smart lighting. From 2020 to 2022, he was entrusted to become Vice Director of the Advanced and Creative Networks Research Center (Ad-CNet RC)-at Telkom University. He is pursuing a doctoral degree at Telkom University, continuing his research about smart lighting. He can be contacted at email: ajigps@telkomuniversity.ac.id.



Roni Habibi    was born in Ciamis, West Java in December 1978. He obtained his Bachelor of Informatics Engineering degree from Nasional University and Master of Informatics degree from Bandung Institute of Technology, Bandung, in 2012 and 2014, respectively. Currently, he is pursuing his Doctoral Program at Indonesian Education University, Bandung. He is involved in research in the field of Information Technology, and IT Risk Management. He is also a lecturer at the Universitas Logistik Dan Bisnis Internasional (ULBI), Bandung. He can be contacted at email: roni.habibi@ulbi.ac.id.



Syafrianita    is a lecturer at the Department of Transportation Management, Universitas Logistik Dan Bisnis Internasional. She has an interest in research using fuzzy numbers in selecting the location of bonded logistics centers. She earned her Ph.D. degree in 2023 at Institut Teknologi Bandung in the field of Transportation. Master's degree received from Department of Transportation, Institut Teknologi Bandung. Bachelor's degree in 2000 at Universitas Pasundan Bandung majoring in Industrial Engineering. She can be contacted at email: syafrianita@ulbi.ac.id.