MDI & PI XGBoost Regional Best Pricing Prediction for Logistics Service

By Agus Purnomo

WORD COUNT	5735	TIME SUBMITTED	30-MAR-2024 11:05AM
		PAPER ID	107933653

MDI and PI XGBoost Regression-Based Methods: Regional Best Pricing Prediction for Logistics Services

Agus Purnomo[1], Aji Gautama Putrada[2], Roni Habibi[3], Syafrianita[4]
[1,4]Faculty of Logistics, Technology and Business, Universitas Logistik Dan Bisnis Internasional, Bandung, Indonesia.
[2]Advanced and Creative, Networks Research Center, Telkom University, Bandung, Indonesia.
[3]Vocational Schools, Universitas Logistik Dan Bisnis Internasional, Bandung, Indonesia.

Article Info

Article history:

Received month dd, yyyy
Revised month dd, yyyy
Accepted month dd, yyyy

Keywords:

Explainable Artificial Intelligence
Retail Price Prediction
Extreme Gradient Boosting
Mean Decrease in Impurity
Permutation Importance

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, Random Forest, and Logistics Regression algorithms. This research contributes to the field by employing MDI (Mean Decrease in Impurity) 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.

Corresponding Author:

Agus Purnomo
Faculty of Logistics, Technology and Business, Universitas Logistik Dan Bisnis Internasional
Jl. Sari Asih No.54, Bandung, Jawa Barat, 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 et al. [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 et al. [5] used factors such as cost differences, level of competition, customer preferences, and local policies, which influenced price adjustments made in each region. However, Riza et al. [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. 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. [10] 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 [11]. Fumagalli et al. [12] showed that Permutation Importance (PI) is another useful method for 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 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:

1. 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.
2. 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 methodology. 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. [10] performed price prediction using regression models and SVM, with a high degree of accuracy. In addition, machine learning approaches are also effective for price comparison. In 2020, Derdouri et al. [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 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 methodology 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.

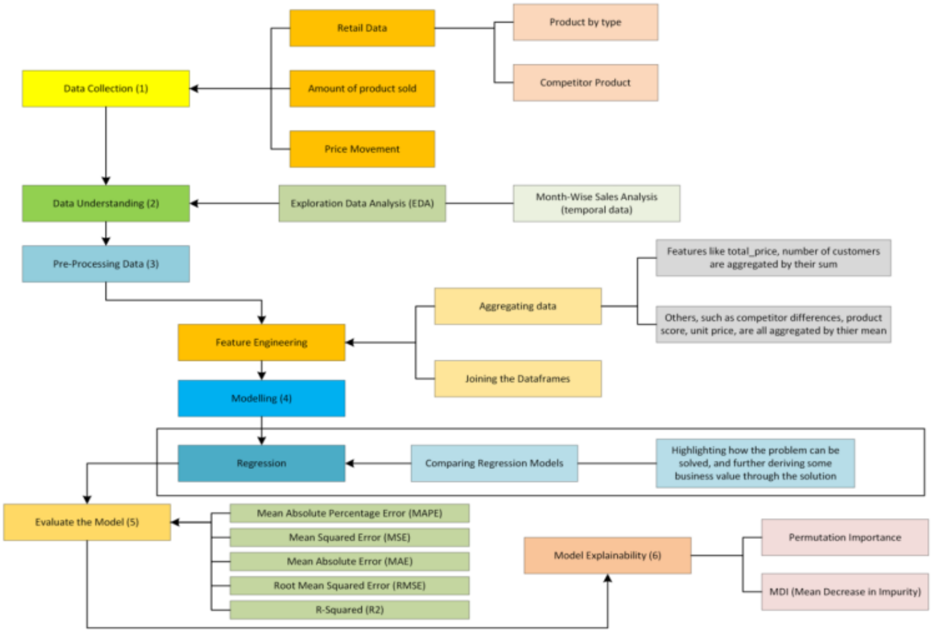
Retail Data
Product by type
Competitor Product
Data Collection (1)
Amount of product sold
Price Movement
Data Understanding (2)
Exploration Data Analysis (EDA)
Month-Wise Sales Analysis (temporal data)
Pre-Processing Data (3)
Features like total_price, number of customers are aggregated by their sum
Aggregating data
Others, such as competitor differences, product score, unit price, are all aggregated by thier mean
Feature Engineering
Joining the Dataframes
Modelling (4)
Regression
Comparing Regression Models
Highlighting how the problem can be solved, and further deriving some business value through the solution
Evaluate the Model (5)
Mean Absolute Percentage Error (MAPE)
Mean Squared Error (MSE)
Mean Absolute Error (MAE)
Root Mean Squared Error (RMSE)
R-Squared (R2)
Model Explainability (6)
Permutation Importance
MDI (Mean Decrease in Impurity)

Figure 1. Proposed methodology

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 is a critical step in scientific research, business analysis, product development, project management, and many other fields where data is needed to make informational, evidence-based decisions.

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 follows:

$$\mu_V = \frac{\sum_{n\in N} V_n}{N} \tag{1}$$

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 LR, RR, Lasso, RFR, GBR, ABR, KNR, and SVR. Regression is a statistical technique to analyze the relationship between two or more variables. Linear Regression (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 (Ridge Regression)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]. Random Forest Regression (RFR) is a method in machine learning that is used to carry out regression or predictions [22]. Gradient Boosting Regression (GBR) is an ensemble algorithm in machine learning used for regression modeling [23]. AdaBoost Regression (ABR) is an ensemble algorithm in machine learning used for regression modeling [24]. K-Nearest Neighbor Regression (KNR) 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. Super Vector 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 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 (Mean Decrease in Impurity) and Permutation Importance (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:

$$Gini(t) = 1 - \sum_{i=1}^{K} (\mathrm{p(i\backslash t)})^2 \qquad (2)$$

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:

$$MDI(X) = \frac{1}{n} \sum_{t} Gini(t) \qquad (3)$$

A higher MDI indicates that the feature is more crucial in making predictions within the Random Forest model. Permutation Importance 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 Permutation Importance (PI) for a feature X can be computed using the formula:

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

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.

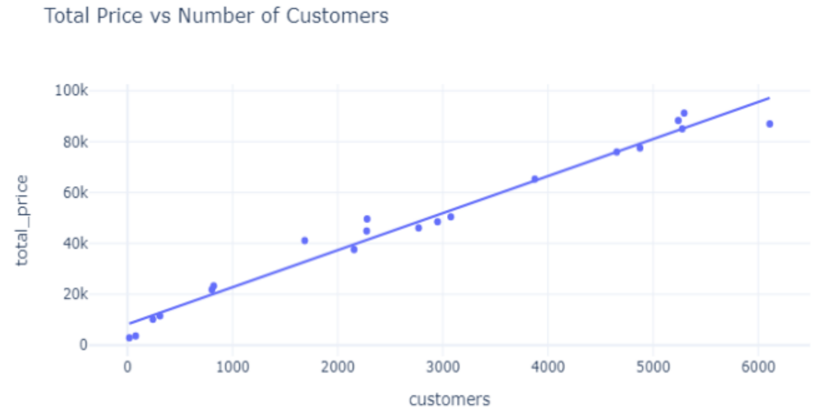

Figure 2. Linear regression and customer per-month analysis; Total price vs customers

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.

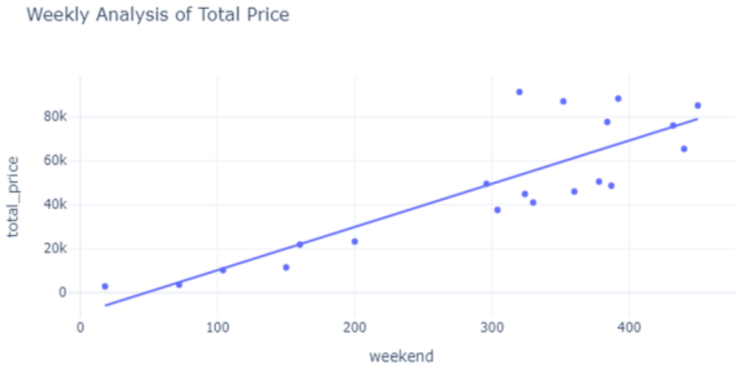

Figure 3. Linear regression and customer per-month analysis; Weakly analysis of total

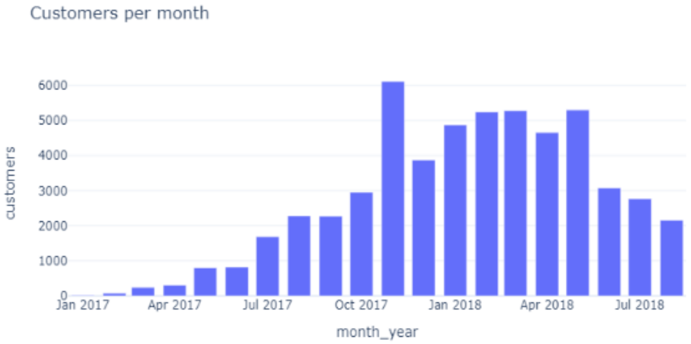

Figure 4. Linear regression and customer per-month analysis; Customer per-month

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.

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.

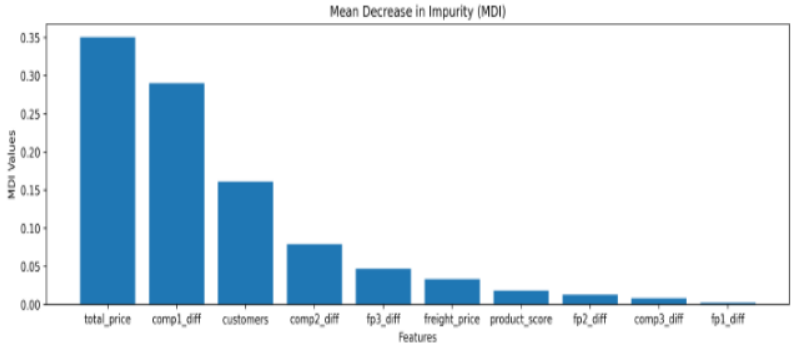

Figure 5. MDI Summary Plot

The Permutation Importance (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.

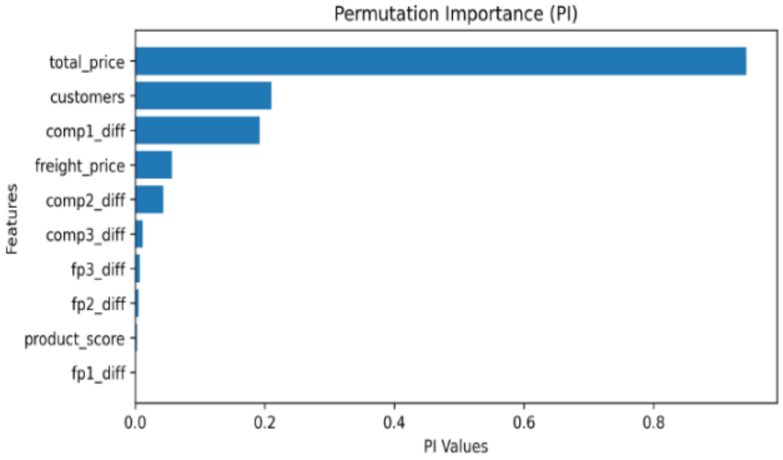

Figure 6. PI Summary Plot

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 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 et al. [35] proposed ABR for house price prediction. Finally, Bonamutial et al. [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 Mean Decrease in Impurity (MDI) and Permutation Importance (PI) techniques in analyzing features. These interpretations, when combined, contribute to a more comprehensive Explainable Artificial Intelligence (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	Permutation Importance
Shahrel et al. [34]	SVR	0.6302	✖	✖
Durganjali et al. [35]	ABR	0.9645	✖	✖
Bonamutial et al. [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 XGBoost Regressor (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 eXplainable Artificial Intelligence (XAI) methods, namely Mean Decrease in Impurity (MDI) and Permutation Importance (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,” *J Mark*, vol. 28, no. 3, pp. 34–38, 1964.

[2] J. Lintner, “Dividends, earnings, leverage, stock prices and the supply of capital to corporations,” *Rev Econ Stat*, pp. 243–269, 1962.

[3] D. Rickert, J. P. Schain, and J. Stiebale, “Local market structure and consumer prices: Evidence from a retail merger,” *J Ind Econ*, vol. 69, no. 3, pp. 692–729, 2021.

[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] F. Riza, “Analisis dan Prediksi Data Penjualan Menggunakan Machine Learning dengan Pendekatan Ilmu Data,” *Data Sciences Indonesia (DSI)*, vol. 1, no. 2, pp. 62–68, 2021.

[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 Int J Geoinf*, vol. 9, no. 2, p. 106, 2020.

[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.

[9] E. Akyildirim, A. Goncu, and A. Sensoy, “Prediction of cryptocurrency returns using machine learning,” *Ann Oper Res*, vol. 297, pp. 3–36, 2021.

[10] M. Jain, H. Rajput, N. Garg, and P. Chawla, “Prediction of house pricing using machine learning with Python,” in *2020 International conference on electronics and sustainable communication systems (ICESC)*, 2020, pp. 570–574.

[11] 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 Behav Soc*, vol. 31, pp. 78–92, 2023.

[12] F. Fumagalli, M. Muschalik, E. Hüllermeier, and B. Hammer, “Incremental permutation feature importance (iPFI): towards online explanations on data streams,” *Mach Learn*, pp. 1–41, 2023.

[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.

[14] S. Gu, B. Kelly, and D. Xiu, “Empirical asset pricing via machine learning,” *Rev Financ Stud*, vol. 33, no. 5, pp. 2223–2273, 2020.

[15] T. Mohd, S. Masrom, and N. Johari, “Machine learning housing price prediction in Petaling Jaya, Selangor, Malaysia,” *Int. J. Recent Technol. Eng*, 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,” *J Clean Prod*, vol. 249, p. 119386, 2020.

[17] L. E. Tomaszewski, J. Zarestky, and E. Gonzalez, “Planning qualitative research: Design and decision making for new researchers,” *Int J Qual Methods*, vol. 19, p. 1609406920967174, 2020.

[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.

[19] F. Fang *et al*., “Cryptocurrency trading: a comprehensive survey,” *Financial Innovation*, vol. 8, no. 1, pp. 1–59, 2022.

[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.

[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,” *Sci Rep*, vol. 11, no. 1, p. 21173, 2021.

[22] E. M. M. der Heide, R. F. Veerkamp, M. L. Van 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,” *J Dairy Sci*, vol. 102, no. 10, pp. 9409–9421, 2019.

[23] S. F. Pane, A. G. Putrada, N. Alamsyah, and M. N. Fauzan, “A PSO-GBR Solution for Association Rule Optimization on Supermarket Sales,” in *2022 Seventh International Conference on Informatics and Computing (ICIC)*, 2022, pp. 1–6.

[24] A. N. Iman, A. G. Putrada, S. Prabowo, and D. Perdana, “Peningkatan Kinerja AMG8833 sebagai Thermocam dengan Metode Regresi AdaBoost untuk Pelaksanaan Protokol COVID-19,” *Jurnal Elektro dan Telekomunikasi Terapan (e-Journal)*, vol. 8, no. 1. pp. 978–985, 2021.

[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.

[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.

[27] M. Abdurohman 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*, 2023.

[28] A. G. Putrada, N. Alamsyah, S. F. Pane, and M. N. Fauzan, “XGBoost for IDS on WSN Cyber Attacks with Imbalanced Data,” in *2022 International Symposium on Electronics and Smart Devices (ISESD)*, 2022, pp. 1–7.

[29] A. Popov, “Feature engineering methods,” in *Advanced Methods in Biomedical Signal Processing and Analysis*, Elsevier, 2023, pp. 1–29.

[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.

[31] D. Lundstrom and M. Razaviyayn, “Distributing Synergy Functions: Unifying Game-Theoretic Interaction Methods for Machine-Learning Explainability,” *arXiv preprint arXiv:2305.03100*, 2023.

[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,” *arXiv preprint arXiv:2109.01433*, 2021.

[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, 2023.

[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.

[35] P. Durganjali and M. V. Pujitha, “House resale price prediction using classification algorithms,” in *2019 International Conference on Smart Structures and Systems (ICSSS)*, 2019, pp. 1–4.

[36] M. Bonamutial and S. Y. Prasetyo, “Exploring the Impact of Feature Data Normalization and Standardization on Regression Models for Smartphone Price Prediction,” in *2023 International Conference on Information Management and Technology (ICIMTech)*, 2023, pp. 294–298.

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 PhD 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
	Dr. (Cand) Aji Gautama Putrada., S.T., M.T. 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, she is pursuing her 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 PhD degree in 2023 at Institut Teknologi Bandung in the field of Transportation. Master's degree received from Department of Transportation, Institut Teknologi Bandung. She can be contacted at email: syafrianita@ulbi.ac.id

MDI & PI XGBoost Regional Best Pricing Prediction for Logistics Service

ORIGINALITY REPORT

6%
SIMILARITY INDEX

PRIMARY SOURCES

1	Min-Joon Kim, Thi-Thu-Huong Le. "Influence Analysis of Real Exchange Rate Fluctuations on Trade Balance Data Using Feature Important Evaluation Methods", Information, 2024 Crossref	38 words — 1%
2	www.researchgate.net Internet	20 words — < 1%
3	Andreas François Vermeulen. "Practical Data Science", Springer Science and Business Media LLC, 2018 Crossref	18 words — < 1%
4	www.indo-intellectual.id Internet	18 words — < 1%
5	wowlic.sookmyung.ac.kr Internet	15 words — < 1%
6	Aji Gautama Putrada, Nur Alamsyah, Syafrial Fachri Pane, Mohamad Nurkamal Fauzan, Doan Perdana. "AUC Maximization for Flood Attack Detection on MQTT with Imbalanced Dataset", 2023 International Conference on Information Technology Research and Innovation (ICITRI), 2023 Crossref	12 words — < 1%

7 Syafrial Fachri Pane, Heriyanto, Aji Gautama Putrada, Nur Alamsyah, Mohamad Nurkamal Fauzan. "The Influence of The COVID-19 Pandemics in Indonesia On Predicting Economic Sectors", 2022 Seventh International Conference on Informatics and Computing (ICIC), 2022
Crossref
12 words — < 1%

8 pure.au.dk
Internet
12 words — < 1%

9 Cheng-Kai Zhang, Rui Zhang, Zhao-Peng Zhu, Xian-Zhi Song, Yin-Ao Su, Gen-Sheng Li, Liang Han. "Bottom hole pressure prediction based on hybrid neural networks and Bayesian optimization", Petroleum Science, 2023
Crossref
11 words — < 1%

10 ejournal.st3telkom.ac.id
Internet
11 words — < 1%

11 orca.cardiff.ac.uk
Internet
11 words — < 1%

12 eprints.utm.my
Internet
10 words — < 1%

13 "Artificial Intelligence, Medical Engineering and Education", IOS Press, 2024
Crossref
9 words — < 1%

14 "Explainable Artificial Intelligence", Springer Science and Business Media LLC, 2023
Crossref
9 words — < 1%

15 Aji Gautama Putrada, Maman Abdurohman, Doan Perdana, Hilal Hudan Nuha. "Machine Learning
9 words — < 1%

Methods in Smart Lighting Toward Achieving User Comfort: A Survey", IEEE Access, 2022
Crossref

16 Helmi Salsabila, Roni Habibi, Nisa Hanum Harani. "Social Media-Based Sentiment Analysis: Electric Vehicle Usage in Indonesia", Indonesian Journal of Computer Science, 2023
Crossref
9 words — < 1%

17 Vivek Singh Rana, Jayanto Mondal, Annu Sharma, Indu Kashyap. "House Price Prediction Using Optimal Regression Techniques", 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2020
Crossref
9 words — < 1%

18 eprints.uad.ac.id
Internet
9 words — < 1%

19 epub.ub.uni-muenchen.de
Internet
9 words — < 1%

20 www.aimspress.com
Internet
9 words — < 1%

21 www.geogsci.com
Internet
9 words — < 1%

22 www.mdpi.com
Internet
9 words — < 1%

23 www.optimization-online.org
Internet
9 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	< 9 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 9 WORDS