Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.152192
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Integrating Bayesian Optimization into Ensemble Logistic Regression for Explainable AI-Based Customer Behavior Analysis

Jeffry[1)*], Azminuddin I. S. Azis [2)], Elisabeth Tri Juliana Kandakon[3)]
[1)2)3)]Institut Teknologi Bacharuddin Jusuf Habibie, Parepare, Indonesia
[1)]jeffry@ith.ac.id, [2)] azminuddinazis@ith.ac.id, [3)] elisabethtrijulianakandakon@gmail.com

Submitted: Aug 12, 2025 | **Accepted**: Sep 3, 2025 | **Published**: Oct 2, 2025

Abstract: Understanding customer behavior is a strategic factor in business decision-making, particularly within the automotive sector, where competition is intense and product variety is diverse. While previous studies often rely on limited demographic variables, such as age and gender, this research advances the field by integrating ensemble logistic regression with Bayesian Optimization for hyperparameter tuning and SHAP-based interpretability. The proposed model incorporates additional features beyond demographics, including vehicle category, product type, vehicle year, dealer branch, and transaction source, to enhance predictive accuracy. The methodology involves data preprocessing through encoding and cleaning, class balancing using SMOTE combined with undersampling, and stratified train-test splitting (80:20). Baseline Logistic Regression achieved an accuracy of 80%, ROC AUC of 0.89, precision of 0.47/0.96, recall of 0.84/0.79, and F1-scores of 0.59/0.89. By applying ensemble logistic regression with Bayesian Optimization, performance improved to 84% accuracy, ROC AUC of 0.92, precision of 0.51/0.98, recall of 0.83/0.84, and F1-scores of 0.63/0.92. SHAP analysis confirmed that the additional features significantly contribute to prediction outcomes. The novelty of this study lies in combining Ensemble Logistic Regression with Bayesian Optimization and SHAP explainability in the automotive domain, offering not only improved accuracy but also interpretability and fairness for business decision-making, providing actionable insights for targeted marketing strategies and product management. Future studies may incorporate broader behavioral and transactional variables to capture more nuanced customer decision patterns..

Keywords: Customer Behavior, Ensemble Logistic Regression, Bayesian Optimization, Explainable AI, SHAP, Automotive Industry

INTRODUCTION

In the midst of the digital transformation era, data has become a strategic asset in business decision-making. Companies that can effectively manage and analyze customer data gain a competitive advantage in responding to market demands (Hariyanti & Kristanti, 2024). The automotive sector in Indonesia, which has experienced rapid growth, is a clear example of how data can be leveraged to understand consumer behavior. According to the Indonesian Automotive Industry Association (GAIKINDO, 2023), the number of vehicles in circulation in Indonesia has reached more than 150 million units, reflecting a significant market potential.

Customer preferences for vehicle transmission types, whether manual or automatic, have been the focus of various international studies. (Manley & Cheng, 2018) reported that demographic variables such as age, gender, and vehicle usage factors play an important role in shaping transmission choices. In addition, geographic location and education level also influence consumer decisions in selecting a vehicle transmission type (Hillel et al., 2021). Historical vehicle purchase data shows a consistent relationship between drivers' demographic attributes and transmission choices, reinforcing the notion that factors such as age and gender are key determinants in this context (Danaher et al., 2019).

Despite the high potential for digitalization, research by (Tama et al., 2021) indicates that most automotive MSMEs (56%) are still in the early stages of Industry 4.0 adoption, and 14% have not even planned for technology

*name of corresponding author

adoption in their strategies. Moreover, infrastructure challenges and limited human resources have contributed to the slow integration of data-driven decision-making on this scale (Hariyanti & Kristanti, 2024).

Machine Learning (ML) offers a solution to these challenges. Logistic Regression (LR) remains a popular choice for binary classification due to its interpretability stemming from probabilistic outputs and transparent log-odds coefficients as well as its computational efficiency on large-scale datasets (Levy & O'Malley, 2020; Saran & Nar, 2025). However, conventional logistic regression is prone to class imbalance issues, where the model tends to be biased toward the majority class and loses performance on non-linear patterns without specific transformation techniques or kernel functions (Chen et al., 2024; Meysami et al., 2023; Saran & Nar, 2025; Zhang et al., 2021).

To address these limitations, ensemble learning approaches have been introduced. Methods such as bagging and boosting have been proven to reduce variance and bias, thereby improving classification accuracy and stability (Ganaie et al., 2022; Jafarzadeh et al., 2021; Ju et al., 2024; Kotsiantis, 2014). In the automotive context, these approaches are more capable of capturing the complexity of customer (Jeffry et al., 2023).

Hyperparameter tuning is essential to further optimize model performance. Bayesian Optimization is an effective parameter search method as it utilizes surrogate functions and probabilistic information (Akiba et al., 2019; Shahriari et al., 2016). This approach is far more efficient than conventional random or grid search methods (Snoek et al., 2012).

Data imbalance is also a concern in this study. The Synthetic Minority Oversampling Technique (SMOTE) combined with Random Undersampling is applied to balance the class distribution. This hybrid approach has been shown to enhance classification model performance in various studies, including applications in the medical and sepsis domains (Liu et al., 2019). Furthermore, research on large observational datasets indicates that combining these methods can significantly improve recall and F1-score (Yang et al., 2024) while supporting adaptive learning models such as ADASYN (He et al., 2008).

Finally, to enhance transparency and interpretability of the model results, the SHAP (SHapley Additive exPlanations) method is employed. SHAP provides a robust explanation of each feature's contribution to the model's decision, based on mathematically consistent Shapley values and model-agnostic properties (Lundberg & Lee, 2017), and has proven effective as a tool for interpretation and feature selection (Mohanty et al., 2024; Mosca et al., 2022; Ponce-Bobadilla et al., 2024; Wang et al., 2024).

This study aims to develop a predictive model for customer preferences regarding vehicle transmission systems using Ensemble Logistic Regression optimized with Bayesian Optimization, complemented by data balancing techniques and SHAP-based model interpretation. Unlike previous studies that predominantly relied on boosting methods such as XGBoost, LightGBM, or CatBoost for customer behavior modeling, this research emphasizes a logistic regression-based ensemble. The reason is that logistic models, when optimized and calibrated, remain highly interpretable and align better with transparency requirements in the automotive industry.

LITERATURE REVIEW

Previous studies have explored various classification approaches in the marketing and automotive domains using Logistic Regression. (Zhang et al., 2021) stated that Logistic Regression is a reliable baseline model for predicting customer preferences. However, its performance tends to decline when applied to datasets with imbalanced class distributions and high-dimensional features (Zhang et al., n.d., 2021).

To address these limitations, ensemble learning has become an increasingly common approach (Cendani & Wibowo, 2022) compared various ensemble methods and concluded that boosting outperforms bagging and stacking. In their experiments, the Gradient Boosting, Extreme Gradient Boosting, and CatBoost models achieved the highest performance on the first dataset, with an accuracy of 81.82%. For the second dataset, Light Gradient Boosting achieved the highest accuracy of 99.25%, while for the third dataset, both Light Gradient Boosting and CatBoost reached a perfect accuracy of 100%.

Similarly, (Lee, 2024) found that ensemble methods such as Bagging, Boosting, and Stacking delivered high accuracy (96–97%) in airline customer satisfaction prediction, with Stacking slightly outperforming the others. (González et al., 2020) also confirmed that Gradient Boosting Machine (GBM) achieved the highest average accuracy (89.51%) across large-scale datasets, outperforming LightGBM (87.77%), XGBoost (86.67%), and CatBoost (79.28%). In HR analytics, (Zohra Sbai, 2025) applied Bayesian Optimization to boosting models, showing that CatBoost achieved 95.8% accuracy and an AUC of 0.98. By combining these approaches, several studies, such as (Jeffry et al., 2023) have successfully increased classification accuracy by up to 10% compared to baseline Logistic Regression.

While ensemble logistic regression and Bayesian Optimization have been applied in various machine learning contexts, their integration with SHAP explainability within the automotive marketing domain remains underexplored. Prior studies on automotive customer segmentation and churn prediction (e.g., (Lee, 2024; Zohra Sbai, 2025)) have primarily used tree-based boosting algorithms. This paper advances the literature by introducing a transparent, interpretable ensemble approach tailored for the automotive sector.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.152192
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Therefore, the present study addresses this gap by integrating Ensemble Logistic Regression with Bayesian Optimization and SHAP-based explainability. Unlike prior works that emphasize boosting algorithms, our approach prioritizes both predictive performance and interpretability, offering a novel framework for understanding and predicting automotive customer preferences.

METHOD

The flowchart in Figure 1 illustrates the research process stages in building and evaluating a classification model using customer data. The process begins with data acquisition, followed by preprocessing steps such as encoding and data cleaning. Next, data balancing is performed using a combination of SMOTE and undersampling to address class imbalance issues. The dataset is then split into training and testing sets with an 80:20 stratified ratio. Two models are developed in parallel: Logistic Regression as the baseline model and a Bagging model optimized using Bayesian Optimization. Each model is evaluated using metrics such as accuracy and ROC AUC, after which the results are interpreted using SHAP to understand the contribution of each feature to the predictions. Finally, a comparative analysis of both models is conducted to determine the best-performing model and its relevance to the research objectives.

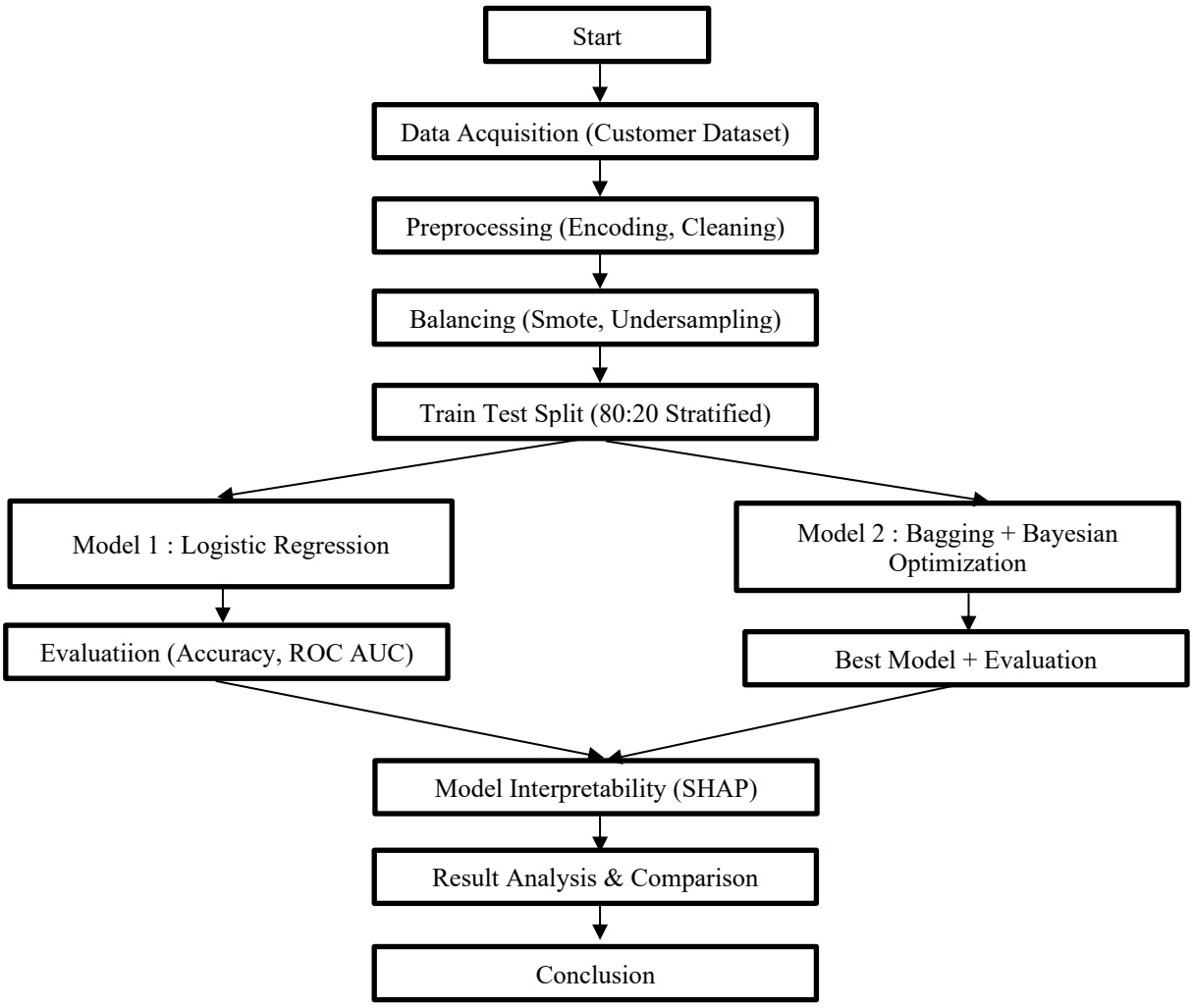

Fig. 1. Research Workflow Diagram

Dataset

The dataset used in this study was obtained from customer transaction records of an automotive dealership located in Makassar, Indonesia, covering the period from 2008 to 2023. The data consists of 221,577 entries and eight main attributes, namely: age, gender, vehicle product, vehicle category, purchase source, dealership branch (Branch), year of purchase, and transmission type as the target label. The target variable was converted into a binary format (1 = automatic, 0 = manual) for classification purposes.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.152192
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 1. Dataset Attributes List

Attribute	Data Type	Description	*Preprocessing*
Gender	Numeric	Customer gender	Passthrough
Age	Numeric	Customer age at the time of transaction	Passthrough
Vehicle Type	Categorical	Name/type of vehicle product	OneHotEncoding
Category	Categorical	Vehicle category (e.g., sedan, SUV, etc.)	OneHotEncoding
Source	Categorical	Purchase source or channel (e.g., new, returning)	OneHotEncoding
Branch	Categorical	Dealership branch where the transaction took place	OneHotEncoding
Year	Numeric	Year of vehicle purchase	Passthrough
Transmission	Categorical	Target variable	Label Encoder

Data Preprocessing

To prepare the data for use in machine learning modeling, preprocessing was carried out as follows: categorical features such as product, category, source, and branch were encoded using the OneHotEncoding technique to convert them into binary numeric features. Numerical features such as age, gender, and year were used without transformation, but were standardized when necessary. The target label transmission was converted into numeric form (0 and 1) using LabelEncoder. This process ensures that all features are in a numerical format compatible with the classification model to be used.

Data Balancing

The dataset exhibits a class imbalance between manual and automatic transmission types. To address this issue, a combined data balancing technique was applied. First, the Random UnderSampler (RUS) was used to reduce the number of samples from the majority class. Then, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples for the minority class. This combination aims to create a more balanced target distribution, thereby reducing bias during model training.

Modeling

This study employs two modeling approaches. The first is Baseline Logistic Regression, which serves as the fundamental classification model and a performance benchmark. The second is Bagging Logistic Regression with Bayesian Optimization, an ensemble model based on the BaggingClassifier with Logistic Regression as the base estimator, combined with hyperparameter optimization using BayesSearchCV.

The models were developed using a scikit-learn pipeline that integrates preprocessing, training, and evaluation. To enhance the predictive performance of the ensemble model, hyperparameter tuning was conducted using the Bayesian Optimization approach. This technique was chosen for its efficiency in exploring the parameter space with fewer iterations compared to conventional methods such as grid search or random search. In this implementation, Bayesian Optimization was executed with 100 iterations and a 5-fold stratified cross-validation strategy. This setup balances computational efficiency and robustness, ensuring that the optimized hyperparameters (such as C, max_depth, and number of estimators) are validated under realistic class imbalance conditions

In this study, the optimized parameters include regularization strength (C), the number of estimators, and additional parameters such as *max_depth* and *max_features*. The optimization process was performed using the Optuna framework, which automatically constructs a Gaussian Process-based surrogate model and selects hyperparameter combinations that minimize the loss function based on an acquisition function, such as Expected Improvement. Tuning was conducted over several iterations for each ensemble scenario. The objective function was defined as the ROC AUC score obtained via 5-fold cross-validation. Bayesian Optimization selects the optimal hyperparameters by:

$$\theta^* = arg\,{}^{max}_{\theta} f(\theta) \qquad (1)$$

where f(θ) represents the ROC AUC score over cross-validation folds.

The search space was defined as follows:

C (regularization strength): [0.01, 10]

Number of estimators (M): [50, 200]

Maximum depth: [3, 10]

These ranges were chosen based on prior studies (Lee, 2024; Zohra Sbai, 2025) and preliminary experiments, to ensure a balance between computational efficiency and model complexity.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.152192
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Pseudocode of the Proposed Algorithm

To ensure the reproducibility of the proposed method, the following pseudocode summarizes the overall workflow of the Ensemble Logistic Regression with Bayesian Optimization framework.

Algorithm: Ensemble Logistic Regression with Bayesian Optimization
1. Preprocess data:
- Handle missing values
- Encode categorical features
- Scale numerical features
- Apply SMOTE & RUS for class balancing
2. Split dataset into Train (80%) and Test (20%) using stratified sampling
3. Initialize Ensemble Logistic Regression (bagging of base classifiers)
4. Apply Bayesian Optimization (Optuna):
- Define search space:
$C \in [0.01, 10]$
$Estimators \in [50, 200]$
$Max\_depth \in [3, 10]$
- Objective: maximize ROC AUC
- Perform 100 trials with 5-fold CV
5. Select best hyperparameters θ^*
6. Retrain Ensemble Logistic Regression with θ^*
7. Evaluate model using Accuracy, Precision, Recall, F1-score, and ROC AUC
8. Apply SHAP for feature importance and interpretability

The pseudocode illustrates the sequential process of our proposed framework, highlighting the integration of ensemble logistic regression, Bayesian optimization, and SHAP-based interpretability to ensure both predictive performance and explainability.

Model Performance Evaluation

The developed models were evaluated using various classification metrics. The primary metrics used include accuracy, precision, recall, and F1-score to assess the balance between positive and negative predictions. In addition, the Area Under the Curve (AUC) from the Receiver Operating Characteristic (ROC) curve and the Precision-Recall Curve (PRC) were employed, as these are particularly important in cases of imbalanced data. Evaluation was conducted on the test data as well as through cross-validation to ensure the consistency and generalizability of the models.

All analysis and experiments were implemented using the Python programming language, utilizing the scikit-learn library for modeling and Optuna for Bayesian optimization.

Accuracy measures the proportion of correct predictions compared to the total number of predictions made by the model, indicating how often the model produces correct outputs.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision indicates how accurate the model's positive predictions are. A high value means that most of the predicted positive instances are indeed correct.

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall measures the model's ability to identify all positive cases in the data. A high value indicates that the model rarely misses positive instances.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

F1-score is the harmonic mean between precision and recall. This metric is particularly useful when the data is imbalanced, as it provides a balance between precision and recall.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

Description:
TP (True Positive): Positive cases that are correctly predicted.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.152192
e-ISSN : 2541-2019
p-ISSN : 2541-044X

TN (True Negative): Negative cases that are correctly predicted.
FP (False Positive): Negative cases that are incorrectly predicted as positive.
FN (False Negative): Positive cases that are incorrectly predicted as negative.

Model Interpretability

To explain how each feature influences the model's predictions, this study employs SHAP (SHapley Additive exPlanations). SHAP is one of the most popular methods within the Explainable Artificial Intelligence (XAI) framework, adopting the Shapley value concept from cooperative game theory. This concept calculates the contribution of each feature by comparing the change in the model's output when the feature is included or excluded from the set of features. In the context of XAI, SHAP helps address the black-box nature of machine learning algorithms, including ensemble models, by providing both quantitative and visual explanations of each feature's influence on the prediction results. This is essential to ensure that the model is not only high-performing but also transparent and trustworthy for decision-makers.

The SHAP analysis process produces two main types of visualizations. The first is the Summary Plot, which displays the distribution of each feature's contribution to predictions across the entire dataset, where the colors indicate feature values (from low to high) and the horizontal axis represents the magnitude of the effect (positive or negative) on the target class probability. The second is the Most Influential Feature Bar Plot, which ranks features based on the average absolute SHAP values, making it easier to identify the dominant factors influencing the model's predictions. By integrating SHAP into this research, the model results are evaluated not only based on accuracy or other performance metrics but also in terms of interpretability. This enables stakeholders to understand the rationale behind the predictions, increases trust in the system, and provides strategic insights that can be leveraged for more effective business decision-making.

RESULT

Model Performance Comparison

This study produced two main models whose performance was compared in classifying customer preferences for vehicle transmission types, namely the baseline single Logistic Regression model and the Ensemble Logistic Regression model using the bagging approach with Bayesian Optimization. The evaluation was conducted on the test data using classification metrics, including accuracy, ROC-AUC, PR-AUC, and SHAP.

Based on the performance evaluation results, both methods—Logistic Regression and Ensemble Logistic Regression with Bayesian Optimization—demonstrated relatively high performance in classifying customer data in terms of distinguishing preferences for vehicle transmission types.

The baseline Logistic Regression model achieved an accuracy of 80%, with a precision score of 0.47 for class 0 and 0.96 for class 1. This indicates that the model performs very well in recognizing class 1, representing customers who tend to choose a particular type of transmission, but shows weaknesses in recognizing class 0, where the precision is relatively low. The recall for class 0 reached 0.84, while the recall for class 1 was 0.79. The F1-score for class 0 was 0.59, and for class 1 was 0.89. The ROC AUC value obtained from this model was 0.8918, indicating a strong capability in distinguishing between the two classes.

In contrast, the proposed model showed slightly better performance than Logistic Regression. It recorded an accuracy of 84%, with a precision of 0.51 for class 0 and 0.98 for class 1. The recall for class 0 was also high at 0.83, and for class 1 at 0.84. The F1-scores achieved were 0.63 for class 0 and 0.92 for class 1.

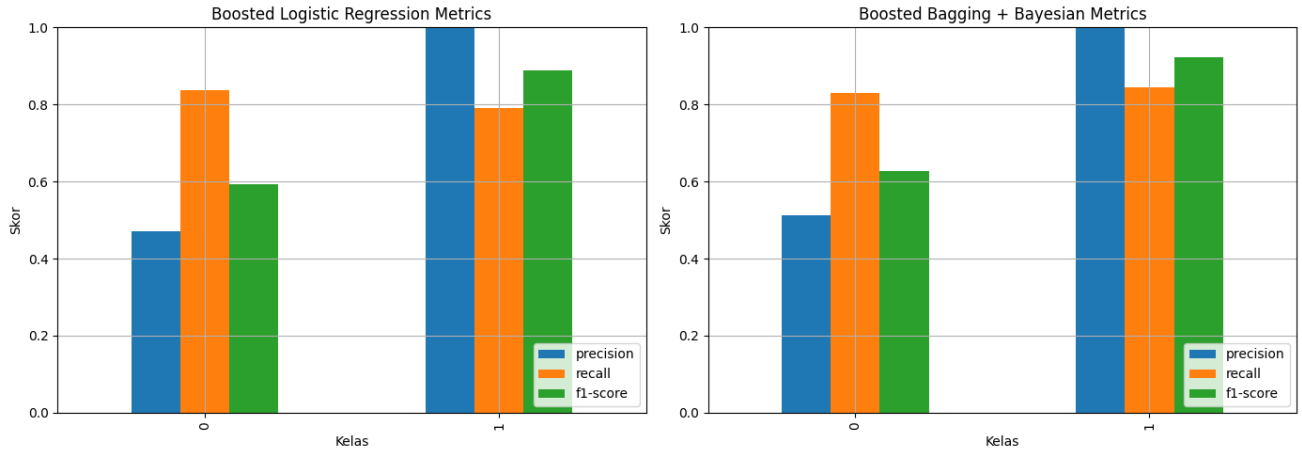

Fig. 2. Model Evaluation Metrics

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.152192
e-ISSN : 2541-2019
p-ISSN : 2541-044X

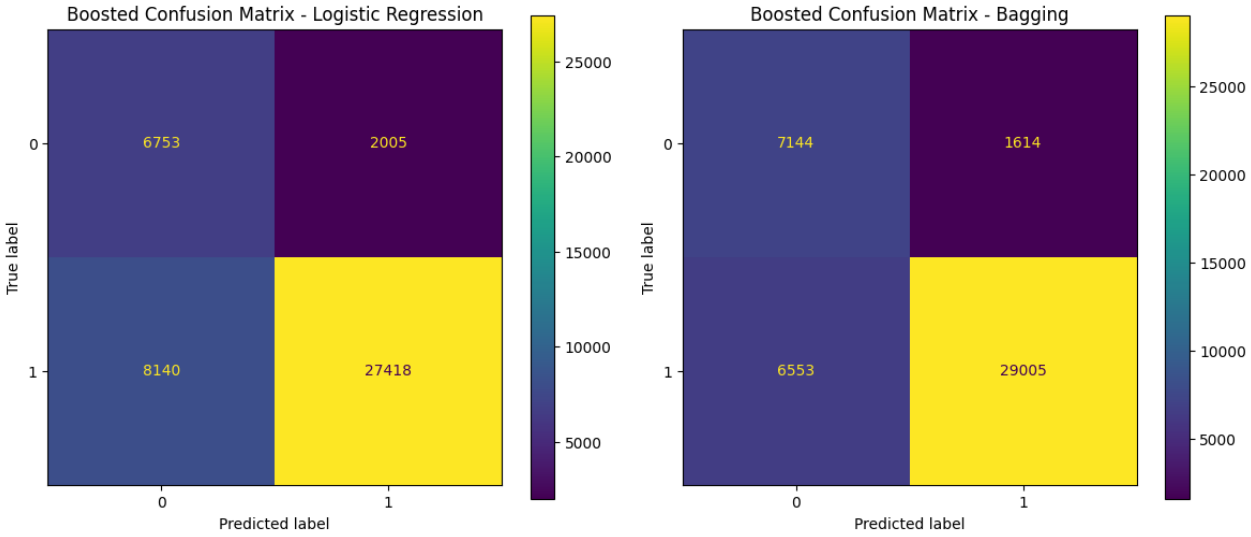

Fig. 3. Confusion Matrix

ROC and Precision-Recall Curve Analysis

The ROC and Precision-Recall (PR) curves were used to evaluate classification quality across all prediction thresholds. Overall, the ROC AUC value of this model reached 0.92, indicating that it has a superior classification capability compared to Logistic Regression. A comparison between the two models shows that, although both methods are capable of classifying the data effectively, the Bagging model outperforms in terms of balancing precision and recall, particularly for the minority class (class 0). This makes the Ensemble Logistic Regression model optimized with Bayesian Optimization more reliable in scenarios that require better detection of the minority class with higher accuracy. Furthermore, the higher F1-score and ROC AUC values reinforce the finding that this model is more effective and stable in classification tasks.

Model Interpretation using SHAP

The SHAP analysis of both the Logistic Regression model and the proposed model (Ensemble Logistic Regression with Bayesian Optimization) provides deeper insights into the contribution of each feature in the prediction process. In the conventional model, the *num_JK* feature emerged as the most influential variable; high values of this feature consistently increased the probability of predicting the positive class, while low values tended to reduce the likelihood of a positive prediction. In addition, categorical features such as *cat_produk1_AVANZA*, *cat_kategori_SUV*, and *cat_produk1_RUSH* also showed significant influence in increasing the likelihood of a positive classification.

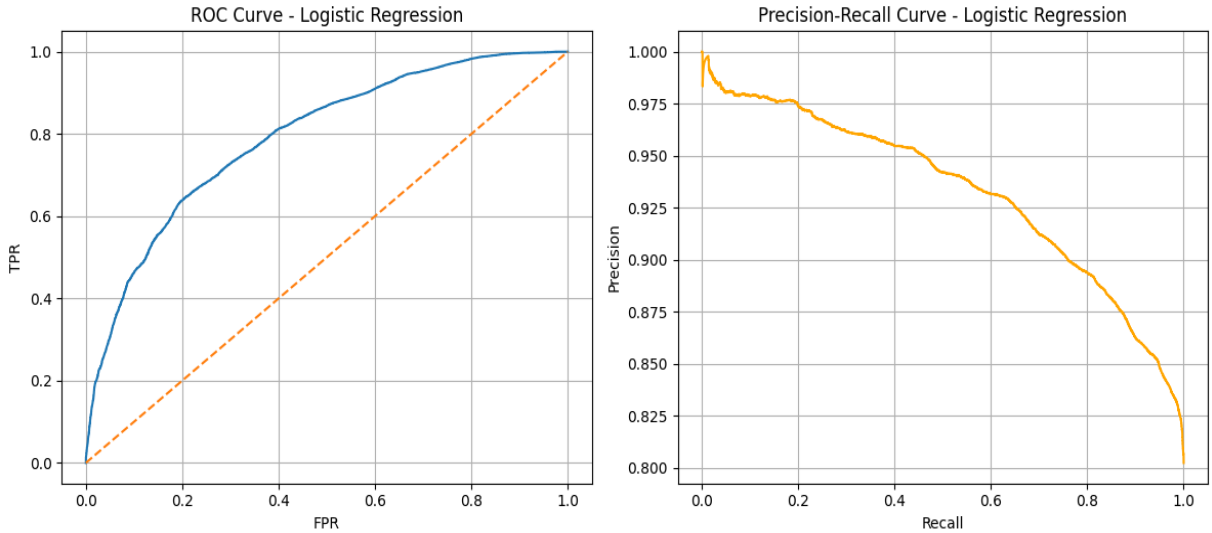

Fig. 4. ROC and PR Curve Visualization of Logistic Regression

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.152192
e-ISSN : 2541-2019
p-ISSN : 2541-044X

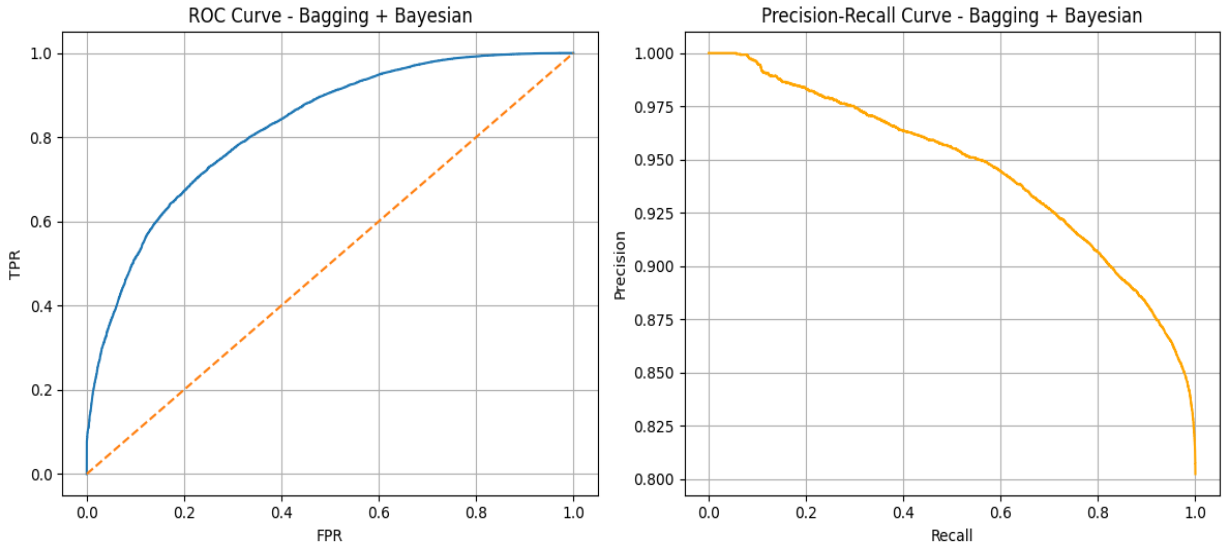

Fig. 5. ROC and PR Curve Visualization of Ensemble Logistic Regression with Bayesian Optimization

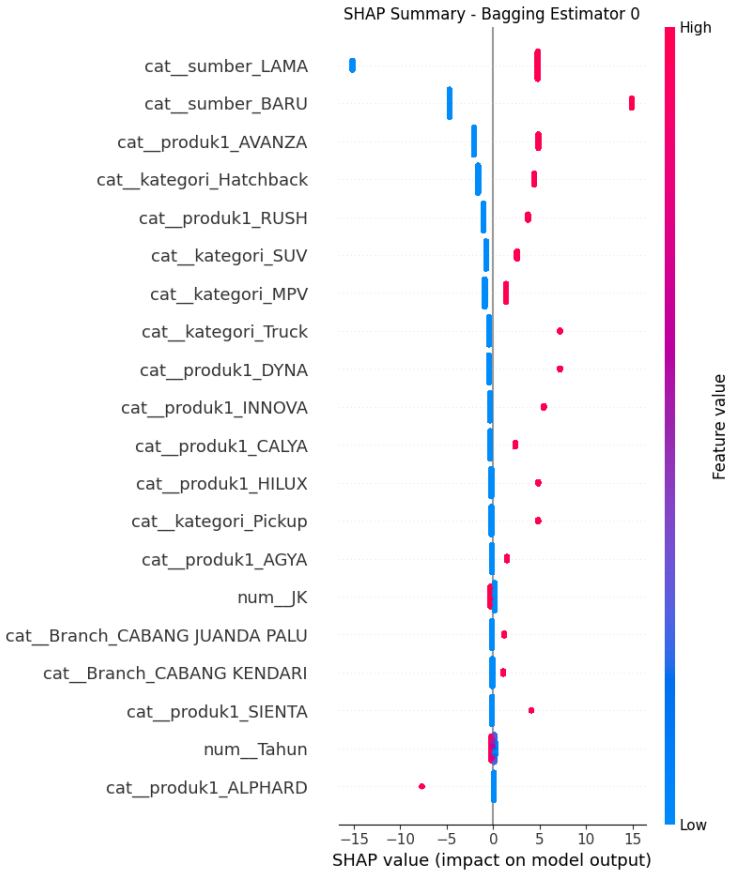

Fig. 6. SHAP Summary for Ensemble Logistic Regression with Bayesian Optimization

One notable finding was related to the data source feature. The *cat_sumber_LAMA* feature, when having high values, contributed strongly and positively to predictions, whereas *cat_sumber_BARU* showed a negative relationship, indicating that the model tended to give preference to data originating from the "LAMA" source. Other features, such as dealer branch and certain product categories, also exhibited positive contributions. Conversely, the *num_Tahun* feature had a negative relationship, where older year values were associated with an increased likelihood of a positive prediction

In contrast, in the proposed model, both *cat_sumber_LAMA* and *cat_sumber_BARU* contributed positively to positive class predictions, differing from the pattern observed in the conventional model. Product-related features

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 9, Issue 4, October 2025
DOI : https://doi.org/10.33395/sinkron.v9i4.152192
e-ISSN : 2541-2019
p-ISSN : 2541-044X

such as *cat_produk1_AVANZA*, *cat_kategori_Hatchback*, and cat_produk1_RUSH remained important within the model structure, consistently contributing to an increased probability of positive predictions.

Although *num_JK* was still relevant in the proposed model, its influence was relatively lower compared to its role in the conventional model. Meanwhile, the contribution pattern of *num_Tahun* remained consistent, with older years showing a positive relationship with the prediction outcome. Overall, the top ten features in the proposed model were dominated by data source attributes and product or vehicle category features, reinforcing the importance of these factors in the ensemble-based model.

In conclusion, both models leverage information from source attributes and product categories to generate predictions, but they differ in how each processes these features. Logistic Regression showed high sensitivity to the *num_JK* feature and made a clear distinction between *cat_sumber_LAMA* and *cat_sumber_BARU*. In contrast, the Ensemble Logistic Regression with Bayesian Optimization tended to assign more balanced weights between the two source types and was less dependent on *num_JK*. This difference is believed to be one of the factors contributing to the superior performance of the proposed model, as reflected in the previous evaluation metrics.

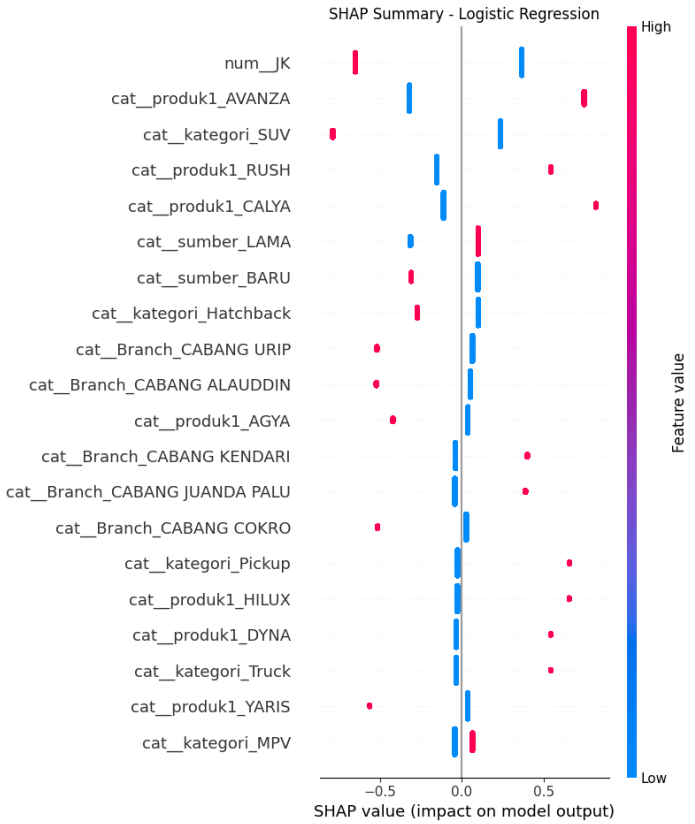

Fig. 7. SHAP Summary for Logistic Regression

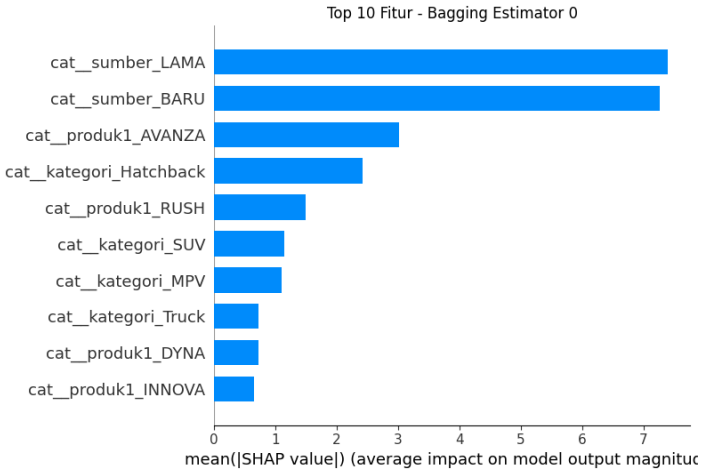

Fig. 8. Top 10 SHAP Features for Ensemble Logistic Regression with Bayesian Optimization

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

DISCUSSIONS

The results of this study demonstrate that the Ensemble Logistic Regression optimized using Bayesian Optimization outperforms the conventional Logistic Regression model. Accuracy improved from 80% to 84%, while the ROC AUC increased from 0.89 to 0.92. The 4% improvement in ROC AUC (from 0.89 to 0.92) translates into a meaningful business implication: fewer misclassified customers in high-cost automotive campaigns, thus reducing wasted marketing resources. Compared to studies such as (Lee, 2024) on automotive customer segmentation and (Zohra Sbai, 2025) on purchase prediction, which relied on boosting algorithms, our approach demonstrates that interpretable logistic-based ensembles can deliver comparable accuracy while maintaining transparency. This aligns with the bias–variance tradeoff theory, where ensemble logistic regression provides stability without sacrificing interpretability. This improvement indicates that the ensemble approach combined with hyperparameter optimization enhances the model's ability to distinguish customer preferences for vehicle transmission types. Such gains are consistent with the theoretical bias-variance tradeoff, where ensemble methods reduce variance and improve generalization compared to a single logistic regression model.

Compared to the baseline, the proposed approach exhibited a better balance between precision and recall, particularly for the minority class (manual transmission). This aligns with previous findings that data balancing techniques such as SMOTE and Random Undersampling can effectively handle class imbalance without compromising overall performance. The ability of the model to capture minority-class patterns is especially important in marketing applications, where niche customer preferences may have significant strategic value.

The SHAP analysis revealed differences in feature contribution patterns between the two models. The baseline model was more sensitive to the gender variable, while the ensemble-based approach assigned more balanced weights to vehicle category, transaction source, and brand. This shift highlights the importance of non-demographic variables, supporting the principles of Explainable AI (XAI) in ensuring interpretability and fairness in predictive modeling. These insights strengthen the practical contribution of the study, as they suggest that automotive customer behavior is better explained by contextual factors rather than demographic ones.

From a practical perspective, these findings can guide the automotive industry in designing more targeted marketing campaigns, optimizing inventory management, and aligning product offerings with market demand. Nevertheless, while the observed 4% increase in accuracy may appear modest, its significance lies in improving decision-making reliability in real-world business contexts. Future research should validate these results on larger and more diverse datasets and explore comparisons with advanced algorithms such as XGBoost or LightGBM to further establish the robustness of the approach.

CONCLUSION

This study demonstrates that Ensemble Logistic Regression with Bayesian Optimization significantly improves predictive performance over conventional Logistic Regression, raising accuracy from 80% to 84% and ROC AUC from 0.89 to 0.92. SHAP-based interpretability identified vehicle category, transaction source, and vehicle brand as the most influential factors, providing actionable insights for targeted marketing and inventory optimization in the automotive industry.

Future research should extend the dataset across multiple regions and time periods, benchmark performance against advanced ensemble methods such as XGBoost and LightGBM, and incorporate customer digital behavior variables. Future research should also evaluate fairness metrics (e.g., equal opportunity, disparate impact) to ensure responsible AI in marketing. Another direction is the integration of customer digital footprint data (such as online browsing and social media activity) and testing deep learning architectures for real-time automotive marketing decision support. These directions are expected to enhance predictive accuracy, broaden applicability, and support real-time decision-making in automotive marketing.

REFERENCES

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Cendani, L. M., & Wibowo, A. (2022). Perbandingan Metode Ensemble Learning pada Klasifikasi Penyakit Diabetes. *Jurnal Masyarakat Informatika*, *13*(1), 33–44. https://doi.org/10.14710/jmasif.13.1.42912

Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review*, *57*(6), 137. https://doi.org/10.1007/s10462-024-10759-6

Danaher, D., Neale, W., McDonough, S., & Donaldson, D. (2019, April 2). *Low Speed Override of Passenger Vehicles with Heavy Trucks*. WCX SAE World Congress Experience. https://doi.org/10.4271/2019-01-0430

GAIKINDO, G. (2023). *Jumlah Kendaraan di Indonesia 147 Juta Unit, 60 Persen di Pulau Jawa – GAIKINDO*. https://www.gaikindo.or.id/jumlah-kendaraan-di-indonesia-147-juta-unit-60-persen-di-pulau-jawa/

*name of corresponding author

Ganaie, M. A., Hu, M., Malik, A. K., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, 105151. https://doi.org/10.1016/j.engappai.2022.105151

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, *64*, 205–237. https://doi.org/10.1016/j.inffus.2020.07.007

Hariyanti, S., & Kristanti, D. (2024). Digital Transformation in MSMEs: An Overview of Challenges and Opportunities in Adopting Digital Technology. *Jurnal Manajemen Bisnis, Akuntansi Dan Keuangan*, *3*(1), Article 1. https://doi.org/10.55927/jambak.v3i1.8766

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

Hillel, T., Bierlaire, M., Elshafie, M. Z. E. B., & Jin, Y. (2021). A systematic review of machine learning classification methodologies for modelling passenger mode choice. *Journal of Choice Modelling*, *38*, 100221. https://doi.org/10.1016/j.jocm.2020.100221

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., & Homayouni, S. (2021). Bagging and Boosting Ensemble Classifiers for Classification of Multispectral, Hyperspectral and PolSAR Data: A Comparative Evaluation. *Remote Sensing*, *13*(21), Article 21. https://doi.org/10.3390/rs13214405

Jeffry, J., Usman, S., & Aziz, F. (2023). Analisis Perilaku Pelanggan menggunakan Metode Ensemble Logistic Regression. *JURNAL TEKNOLOGI DAN ILMU KOMPUTER PRIMA (JUTIKOMP)*, *6*(2), 90–97.

Ju, J., Lee, E., & Park, S. (2024). Comparative Analysis of Ensemble Machine Learning Models for Personalized In-Vehicle Infotainment Recommendation Systems. *Adjunct Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 45–50. https://doi.org/10.1145/3641308.3685021

Kotsiantis, S. B. (2014). Bagging and boosting variants for handling classifications problems: A survey. *The Knowledge Engineering Review*, *29*(1), 78–100. https://doi.org/10.1017/S0269888913000313

Lee, M. (2024). COMPARISON OF BAGGING, BOOSTING, AND STACKING ENSEMBLE MODELS FOR AIRLINE CUSTOMER SATISFACTION ANALYSIS. *FaST - Jurnal Sains Dan Teknologi (Journal of Science and Technology)*, *8*(1), Article 1. https://doi.org/10.19166/jstfast.v8i1.8166

Levy, J. J., & O'Malley, A. J. (2020). Don't dismiss logistic regression: The case for sensible extraction of interactions in the era of machine learning. *BMC Medical Research Methodology*, *20*(1), 171. https://doi.org/10.1186/s12874-020-01046-3

Liu, S., Lun Ong, M., Kin Mun, K., Yao, J., & Motani, M. (2019, December 30). *Early Prediction of Sepsis via SMOTE Upsampling and Mutual Information Based Downsampling*. 2019 Computing in Cardiology Conference. https://doi.org/10.22489/CinC.2019.239

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Manley, E., & Cheng, T. (2018). Exploring the role of spatial cognition in predicting urban traffic flow through agent-based modelling. *Transportation Research Part A: Policy and Practice*, *109*, 14–23. https://doi.org/10.1016/j.tra.2018.01.020

Meysami, M., Kumar, V., Pugh, M., Lowery, S. T., Sur, S., Mondal, S., & Greene, J. M. (2023). Utilizing logistic regression to compare risk factors in disease modeling with imbalanced data: A case study in vitamin D and cancer incidence. *Frontiers in Oncology*, *13*, 1227842. https://doi.org/10.3389/fonc.2023.1227842

Mohanty, P. K., Francis, S. A. J., Barik, R. K., Roy, D. S., & Saikia, M. J. (2024). Leveraging Shapley Additive Explanations for Feature Selection in Ensemble Models for Diabetes Prediction. *Bioengineering*, *11*(12), 1215. https://doi.org/10.3390/bioengineering11121215

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). SHAP-Based Explanation Methods: A Review for NLP Interpretability. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 4593–4603). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.406/

Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, *17*(11), e70056. https://doi.org/10.1111/cts.70056

*name of corresponding author

Saran, N. A., & Nar, F. (2025). Fast binary logistic regression. *PeerJ Computer Science*, *11*, e2579. https://doi.org/10.7717/peerj-cs.2579

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, *104*(1), 148–175. https://doi.org/10.1109/JPROC.2015.2494218

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*, *25*. https://papers.nips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

Tama, I. P., Tantrika, C. F. M., Hardiningtyas, D., & Mohamad, E. (2021). Review of Industry 4.0 Strategy and Organization Readiness Level of Automotive SMEâ€™s in Indonesia. *APMBA (Asia Pacific Management and Business Application)*, *9*(3), 313–324. https://doi.org/10.21776/ub.apmba.2021.009.03.9

Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1), 44. https://doi.org/10.1186/s40537-024-00905-w

Yang, C., Fridgeirsson, E. A., Kors, J. A., Reps, J. M., & Rijnbeek, P. R. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*, *11*(1), 7. https://doi.org/10.1186/s40537-023-00857-7

Zhang, L., Geisler, T., Ray, H., & Xie, Y. (2021). Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function. *Journal of Applied Statistics*, *49*(13), 3257–3277. https://doi.org/10.1080/02664763.2021.1939662

Zhang, L., Ray, H., Priestley, J., & Tan, S. (n.d.). A descriptive study of variable discretization and cost-sensitive logistic regression on imbalanced credit data. *Journal of Applied Statistics*, *47*(3), 568–581. https://doi.org/10.1080/02664763.2019.1643829

Zohra Sbai, T. V. (2025). Bayesian Optimized Boosted Ensemble models for HR Analytics—Adopting Green Human Resource Management Practices. *International Journal of Technology*, *16*(2), 291–319. https://doi.org/10.14716/ijtech.v16i2.7277

*name of corresponding author