BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 2281-2298
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp2281-2298

LEVERAGING XGBOOST, LIGHTGBM, AND CATBOOST FOR ENHANCED CUSTOMER SEGMENTATION IN THE AUTOMOTIVE INDUSTRY

Novri Suhermi[1]*, Rahida Rihhadatul Aisy[2], Aulia Afifatur Rohmah[3], Anis Alif Nurhayati[4], Agnes Nathania Pramesty[5], Aura Lovi Ardanika[6], Fauziyah Nurul Isnaini[7]

[1,2,3,4,5,6,7]*Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember*
Jln. Teknik Kimia, Keputih, Kec. Sukolilo, Surabaya, 60111, Indonesia

*Corresponding author's e-mail: * **novri.suhermi@its.ac.id***

Article Info

Article History:

Received: 15th August 2025
Revised: 1st December 2025
Accepted: 10th March 2026
Available online: 8th April 2026

Keywords:

CatBoost;
Automotive industry;
Customer segmentation;
XGBoost;
LightGBM.

ABSTRACT

This study evaluates the performance of three gradient boosting algorithms, XGBoost, LightGBM, and CatBoost, for customer segmentation in the automotive industry. Utilizing a dataset of 8,068 training and 2,627 testing observations with 11 demographic and behavioral variables, the research aims to classify customers into four segments. The methodology includes preprocessing (handling missing values, encoding), hyperparameter tuning via Randomized Search Cross-Validation, and evaluation using ROC AUC. Results indicate that XGBoost outperforms other models, achieving an AUC of 0.5837 on testing data with significant variables, while LightGBM and CatBoost scored 0.5834 and 0.5759, respectively. Key findings highlight the importance of feature selection, with Age, Profession, and Spending Score being the most influential predictors. The study concludes that XGBoost is the most robust for segmentation tasks, though all models exhibit challenges in distinguishing overlapping classes. These insights can guide data-driven marketing strategies in automotive and related sectors.

How to cite this article:

N. Suhermi, *et al.*, "LEVERAGING XGBOOST, LIGHTGBM, AND CATBOOST FOR ENHANCED CUSTOMER SEGMENTATION IN THE AUTOMOTIVE INDUSTRY," *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 2281-2298, Sep, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

In an era of globalization and increasingly fierce industrial competition, companies are required to continuously develop innovative and data-driven business strategies in order to maintain competitiveness and expand market share. One approach that has proven effective in improving marketing and sales efficiency is a customer segmentation strategy. Customer segmentation enables companies to divide a large and heterogeneous market into more homogeneous consumer groups with similar needs and preferences. This strategy not only helps companies target customers more accurately but also allocates marketing resources more efficiently [1].

The automotive company that is the subject of this study has successfully implemented a customer segmentation approach in its existing market. In practice, the company classifies all its customers into four main segments based on specific attributes and behaviors. After segmentation is completed, the company's marketing team implements different communication approaches and sales strategies for each segment. This personalized approach has proven to have a significant positive impact on the effectiveness of marketing campaigns and increased sales figures [2].

Building on its success in the domestic market, the company now plans to expand its market reach into new regions while continuing to rely on its existing product lines, namely P1, P2, P3, P4, and P5. Initial market research indicates that the characteristics and behavior of consumers in the new market are similar to those in the previously established market. Therefore, the company intends to apply the same segmentation strategy in its approach to this new market. Currently, the company has successfully identified 2,627 potential new customers who could become the target market in the expansion area [2].

However, to replicate the success of the previous strategy, the company faces a major challenge: grouping these new potential customers into the appropriate segments (A, B, C, or D) using the most suitable method, as was done with existing customers [2]. Given the large volume of data and the importance of accuracy in customer segmentation, a manual approach is inefficient and prone to errors. Therefore, a technology-based solution, particularly a machine learning-based approach, is needed to map new potential customers into the appropriate segments.

In this context, a supervised learning approach is considered most suitable because historical data from existing customers that have been segmented can be used as training data, with segment labels as target variables. By utilizing customer attributes as features and customer segments as labels, supervised learning models can be trained to recognize patterns that distinguish each segment. Once the model is built and validated, this method can be used to predict the segment of new potential customers with a high degree of accuracy.

Three popular gradient boosting decision tree-based algorithms, namely Extreme Gradient Boosting (XGBoost), Light Gradient Boosting (LightGBM), and Categorical Boosting (CatBoost), were considered in this study because they perform well in classification on large datasets with mixed data types, namely numerical and categorical, and can address data imbalance [3]. Previous studies also support the effectiveness of these three methods in the context of customer segmentation. Research conducted by [4] shows that XGBoost can achieve high accuracy in classifying credit card customers. Research by [5] shows that LightGBM performs very well in classifying customers at PT Kasir Pintar Internasional. In addition, research by [6] shows that the CatBoost method demonstrates the best overall performance across all metrics in e-commerce consumer classification.

Based on the background described above, this study aims to classify new potential customers into one of four primary segments, Segment A, B, C, or D, by employing supervised learning algorithms, namely XGBoost, LightGBM, and CatBoost. Furthermore, this study seeks to compare the performance of these three algorithms in handling class imbalance and to determine the optimal model for customer segmentation. Additionally, the research aims to identify the key characteristics of each segment based on the best model and feature importance analysis, thereby providing deeper insights into the profile of each segment. The results of this study are expected to serve as a foundation for formulating more effective, data-driven marketing strategies to support the automotive company's market expansion efforts.

2. RESEARCH METHODS

2.1 Machine Learning Methods

2.1.1 XGBoost

XGBoost (Extreme Gradient Boosting) is one of the ensemble-based machine learning algorithms developed to improve the accuracy and efficiency of the gradient boosting method. It builds a prediction model incrementally through a series of decision trees, where each new tree is added to correct the errors of the previous prediction. XGBoost approaches the gradient boosting framework enhanced using second-order optimization, which optimizes the loss function by considering both the first derivative (gradient) and the second derivative (hessian), allowing for more accurate estimation in each learning iteration [7].

XGBoost also features explicit regularization (L1 and L2) to avoid overfitting, as well as shrinkage or learning rate adjustment to control the impact of each tree on the final model. In addition, XGBoost also supports column subsampling and row subsampling, which not only help reduce overfitting but also speed up the training process. XGBoost has the ability to handle missing values automatically, which is one of the important advantages in its application to real-world data, and has been proven with previous research [8]-[12]. Missing values do not need to be explicitly imputed because during the training process, the algorithm will internally treat the value as its own branch. This branch is chosen based on the best loss reduction, thus maintaining model performance without requiring complex pre-processing of missing values [13]. The XGBoost algorithm can be broken down as follows.

1. Dataset Initialization
   The training dataset consists of pairs (x_k, y_k), where $(x_{k1}, x_{k2}, \dots, x_{kd})$ is the input feature. With y_k is the target, which can be numeric (regression) or binary (classification).
2. Initial Model Initialization
   Initial Model $F_0(x)$ is calculated by minimizing the loss function with respect to the target.
   Regression (squared error):

$$F_0(x) = \frac{1}{n}\sum_{k=1}^{n} y_k \,. \tag{1}$$

   Classification (log-loss):

$$F_0(x) = \log\left(\frac{p}{1-p}\right), \tag{2}$$

   with value $p = \frac{1}{n}\sum_{k=1}^{n} y_k$.
3. Calculate Gradient and Hessian values (second derivative)
   For every k-th data calculate the value of the gradient and Hessian.

   Gradient:

$$g_k^{(t)} = \frac{\partial L(y_k, \widehat{y_k})}{\partial \widehat{y_k}}. \tag{3}$$

   Hessian:

$$h_k^{(t)} = \frac{\partial^2 L(y_k, \widehat{y_k})}{\partial \widehat{y_k^2}}. \tag{4}$$

4. Build a Regression Tree (Base Learner)
   Build a decision tree $h_t(x)$ to map inputs to outputs based on gradient and Hessian information. The tree is built by selecting splits based on maximum Gain.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]-\gamma, \quad (5)$$

with:

G_L, H_L : the sum of the gradient and the Hessian on the left-hand side;
G_R, H_R : the sum of the gradient and the Hessian on the right-hand side;
λ : L2 regulation parameter;
γ : split complexity penalty.

5. Calculate the Output of Each Leaf (Leaf Value)
For each j-th leaf, the predicted value is calculated as follows

$$\omega_j = -\frac{\sum_{k\in I_j} g_k^{(t)}}{\sum_{k\in I_j} h_k^{(t)}+\lambda}, \quad (6)$$

with I_j is the index of the sample that goes to leaf j.

6. Update Model
The model is updated by adding the prediction results from the tree:

$$F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x), \quad (7)$$

with η is the learning rate.

7. Final Model
After the iteration is complete, the final model obtained is as follows.

$$F_T(x) = F_0(x) + \sum_{t=1}^{T} \eta \cdot h_t(x). \quad (8)$$

XGBoost is a powerful algorithm that builds predictive models efficiently and accurately through a gradient-based boosting approach. By utilizing second-order optimization, XGBoost considers both the gradient and the Hessian in the learning process, thus providing more precise estimates at each iteration [14].

2.1.2 LightGBM

LightGBM is one of the Gradient Boosting models that can be used in ranking, classification, regression, and several other machine learning tasks in a fast, distributed, and high-performance manner based on decision tree algorithms. The advantages of the LightGBM method compared to other Gradient Boosted Decision Trees (GBDT) methods are more memory efficiency, better accuracy, and suitable for handling large-scale data [15]. Several works also proved that LightGBM has superior ability across diverse applications [16]-[20].

LightGBM is a Gradient Boosting algorithm that works by building models incrementally (iteratively) using decision trees. Each new model is built to correct the errors of the previous model. The LightGBM algorithm can be detailed as follows:

1. The prediction model iteration process at step t with the following equation

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i), \quad (9)$$

where:
$\hat{y}_i^{(0)} = 0$;
i : index of data in the training dataset ($i = 1,2,\dots,n$);
t : index for the resulting tree ($t = 1,2,\dots,T$);
$f_t(x_i)$: predicted value at iteration t for data i;
$\hat{y}_i^{(t)}$: total prediction result until iteration t for the i-th data.

2. Define the loss function by adding a regularization component, expressed in equation

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i{}^{(t)}\right) + \sum_{t=1}^{T} \Omega(f_t), \tag{10}$$

with the following regularization

$$\Omega(f_t) = \Upsilon T + \frac{1}{2}\lambda\|\omega\|^2, \tag{11}$$

where:

$\mathcal{L}^{(t)}$: loss function at iteration t;
ω : prediction results from the tree;
Υ, λ : regularization parameters;
T : number of trees.

3. Calculate the gradient or derivative of $\mathcal{L}^{(t)}$ with respect to $\hat{y}_i$ for each data point.
4. Select $a\%$ of data with the largest gradient.
5. Select $b\%$ of the remaining data from step 4 (small gradient) randomly.
6. Multiply the small gradient data by the adjustment factor $\frac{1-a}{b}$.
7. Build a tree using the combined data from steps 4 and 6.
8. Return to step 1 until iteration T [21].

2.1.3 CatBoost

CatBoost (Categorical Boosting) is an open-source implementation of the Gradient Boosted Decision Trees (GBDT) algorithm that is specifically developed to handle various supervised machine learning tasks, such as classification and regression [22]. Gradient Boosting is an ensemble technique that aims to optimize the loss function by forming a final model that combines several individual models [23]. The main difference between CatBoost and other Gradient Boosting methods, such as XGBoost and LightGBM, is the use of symmetric and balanced trees (oblivious trees) as the base predictor to achieve optimal prediction speed. In oblivious trees, the tree is built symmetrically, where each split at a level uses the same criteria for all nodes at that level, so the tree grows symmetrically and balanced. This approach provides several advantages for CatBoost, such as shorter training and prediction times, the ability to reduce the risk of overfitting, and improved GPU usage efficiency [24]. Several studies have shown that CatBoost is the best model for various applications [25]-[29].

In applying the basic principles of gradient boosting, CatBoost employs a specialized approach to address categorical features and overfitting. The main steps in the CatBoost algorithm are as follows:

1. Dataset Initialization
   The training dataset consists of pairs (x_k, y_k), where: $x_k = (x_{k1}, x_{k2}, \ldots, x_{km})$ is input features (consisting of numerical and categorical features) and y_k is the target, and can be numerical (regression) or binary (classification).

2. Transformation of Categorical Features (Ordered Target Statistics)
   Create a random permutation σ from the training dataset, then for each categorical feature x_{ik} and k-th example, the categorical value is transformed into a numerical value $\hat{x}_{ik}$ using the following formula:

$$\hat{x}_{ik} = \frac{\sum_{j:\sigma(j)<\sigma(k)}^{n} 1_{\left\{x_j^i = x_k^i\right\}} \cdot y_j + ap}{\sum_{j:\sigma(j)<\sigma(k)}^{n} 1_{\left\{x_j^i = x_k^i\right\}} + a}, \tag{12}$$

where:

$a > 0$: Smoothing parameter;
p : average target of all data (prior);
$\hat{x}_{ik}$: the new numerical value of the i-th categorical feature for the k-th sample;
$1_{\left\{x_j^i = x_k^i\right\}}$: Indicator function (1 if equal, 0 otherwise).

3. Model Initialization
The initial model $F_0(x)$ is initialized by minimizing the loss function against the target value as follows:

$$F_0(x) = \arg\min_c \sum_{i=1}^{n} L(y_i, c)\,. \tag{13}$$

Examples:

Regression (squared error) : $F_0(x) = \bar{y}$.
Classification (log-loss) : $F_0(x) = \log\left(\frac{p}{1-p}\right)$, with $p = \frac{1}{n}\sum y_k$.

4. Gradient Boosting Iteration (Ordered Boosting)
   a. Calculating Negative Gradients
   For each data point *k*, calculate the partial derivative of the following loss function:

$$g_t(x_k, y_k) := \left.\frac{\partial L(y_k, s)}{\partial s}\right|_{s=F_{t-1}(x_k)} \tag{14}$$

   b. Building a Base Learner Model
   Find a decision tree function $h_t(x) \in \mathcal{H}$ that minimizes the following value:

$$h_t = arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{k=1}^{n} (g_t(x_k) - h(x_k)^2 \tag{15}$$

   CatBoost uses oblivious trees (symmetric trees), where all nodes at the same level use the same split.

   c. Updating the Model
   The model is as follows:

$$F_t(x) = F_{t-1}(x) + \eta \cdot h_t(x) \tag{16}$$

   Where η is the learning rate.

5. Final Model
After iteration T is complete, the final model is obtained:

$$F_T(x) = F_0(x) + \sum_{t=1}^{T} \eta \cdot h_t(x)\,. \tag{17}$$

This model is then used to predict target values in new data.

2.1.4 Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) curve is a tool used to evaluate the performance of a classification model by plotting the True Positive Rate (TPR) or recall against the False Positive Rate (FPR) at various thresholds [24]. TPR is the ratio of correctly detected positive samples to the total number of positive samples, while FPR is the ratio of negative samples incorrectly classified as positive to the total number of negative samples. The ROC curve visualizes the tradeoff between TPR and FPR, thereby helping to analyze the impact of threshold changes on model performance. The Area Under the ROC Curve (AUC) summarizes the performance of the ROC curve into a single scalar value between 0 and 1. An AUC value of 0.5 indicates random classification, while a value close to 1 indicates good discriminative ability [30]. Statistically, AUC reflects the probability that the model will assign a higher score to a random positive sample compared to a random negative sample [31]. ROC AUC is independent of the threshold, making it effective for evaluating models on class-imbalance datasets or for determining the optimal threshold based on the TPR-FPR balance [14]. This metric is very useful for comparing the overall discriminative ability of classification models.

2.2 Dataset and Analysis

2.2.1 Data Sources

The dataset used in this study is secondary data obtained from the *Analytics Vidhya* platform, titled *Customer Segmentation*. It contains customer information from an automotive company and includes various demographic and behavioral attributes. The data consists of 8,068 observations in the training set and 2,627 observations in the testing set, with a total of 11 variables comprising both numerical and categorical data types. Key variables include age, annual income, annual spending, and other relevant attributes for customer segmentation analysis.

2.2.2 Research Variables

A proper understanding of these variables and their respective data types is fundamental to ensuring the accuracy and relevance of the subsequent preprocessing, statistical testing, and predictive modeling. This study uses 11 variables, which are presented in Table 1.

Table 1. Variable in Research

Variable	Scale of Data	Description
ID	Nominal	Unique ID for each customer (no quantitative meaning)
Gender	Nominal	Customer's gender (Male or Female)
Ever Married	Nominal	Customer's marital status (Yes or No)
Age	Ratio	Customer's age in years (has absolute zero and meaningful ratio)
Graduated	Nominal	Customer's graduation status (Yes or No)
Profession	Nominal	Customer's type of profession
Work Experience	Ratio	Customer's work experience in years
Spending Score	Ordinal	Customer's spending score
Family Size	Ratio	Number of family members, including the customer
Var 1	Nominal	Anonymized customer category
Segmentation	Nominal	Customer segment label (A, B, C, or D)

2.2.3 Analysis Steps

The analytical procedure employed in this study is outlined and presented in a flowchart in Fig. 1. The following flowchart is designed to provide a concise and systematic overview of the research flow in customer segmentation using the XGBoost, LightGBM, and CatBoost algorithms. The process begins with data collection, followed by the pre-processing stage, such as handling missing values and encoding data to suit the type of algorithm being used. Once the data is ready, exploratory data analysis (EDA) is conducted to understand the patterns and characteristics of the variables. The next stage is modeling using the three tree-based algorithms, with the results evaluated using specific evaluation metrics such as the ROC AUC score. In addition, a feature importance analysis is carried out to identify the most influential variables in customer segmentation. The process concludes with drawing conclusions and providing recommendations based on the model evaluation results. This flowchart is expected to help readers understand the entire workflow comprehensively.

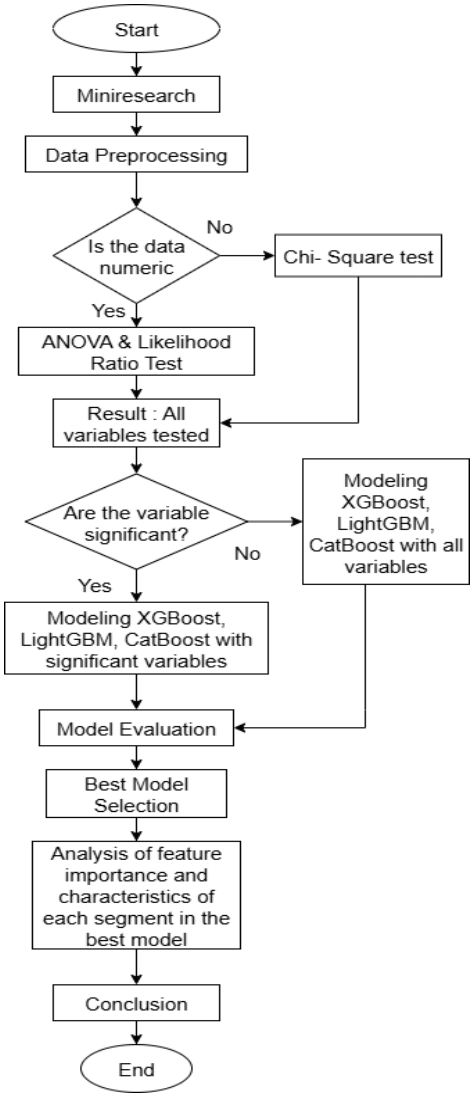

Figure 1. Flowchart Analysis

3. RESULTS AND DISCUSSION

3.1 Pre-processing

The pre-processing phase is conducted to prepare the data before it is used in modeling. At the initial stage of pre-processing, a check was made on the presence of missing values in the dataset. The results show that there are several columns with missing values, including Ever Married (1.90%), Graduated (0.91%), Profession (1.45%), Var 1 (1.22%), Family Size (4.30%), and Work Experience (10.24%). Since all columns with missing values were below the 10% threshold commonly used in data analysis, it was decided to keep the features and impute rather than delete the data.

3.2 Descriptive Statistics

Descriptive statistics were used in this study to figure out the distribution of numerical variables in the dataset, as shown in Table 2.

Table 2. Descriptive Statistics for Numerical Variables

Variable	Minimum	Maximum	Mean
Age	18	89	43.467
Work Experience	0	14	2.473
Family Size	1	9	2.856

To test the relationship between the response variable and its input variables, the Likelihood Ratio Test in the context of multinomial logistic regression was used to find the relationship between the response variable and the numerical input variables. The results of the Likelihood Ratio Test can be seen in Table 3. From Table 3, it is known that the variables Age and Family Size have a significant relationship with the variable Segmentation at a significant level of 5% because they have a p-value < 0.05. Meanwhile, the variable Work Experience does not have a significant relationship with the variable Segmentation at a significant level of 5% because it has a p-value > 0.05.

Table 3. Likelihood Ratio Test (LRT) Results for Numerical Variables

Numeric Input	Degrees of Freedom	LRT Statistics	p-value	Description
Age	3	1298.139	0.000	Significant
Work Experience	3	7.419	0.060	Not Significant
Family Size	3	190.322	0.000	Significant

In this study, the relationship between the response variable and the categorical input variable was analyzed using the Chi-square test. The result from this test is presented in Table. 4, which shows that all categorical input variables have a significant relationship with the Segmentation variable at a significance level of 5% because they have a p-value < 0.05. Based on these results, it can be concluded that all categorical input variables that were statistically tested have a significant relationship with the Segmentation variable, making them relevant for further analysis in the context of customer segmentation.

Table 4. Chi-square Test Results for Categorical Variables

Categoric Input	Degrees of Freedom	Chi-Square	p-value	Description
Gender	3	11.429	0.010	Significant
Ever Married	3	1368.637	0.000	Significant
Graduated	3	1060.340	0.000	Significant
Profession	24	2590.801	0.000	Significant
Spending Score	6	1516.879	0.000	Significant
Var 1	18	242.705	0.000	Significant

3.3 Modeling of All Variables

3.3.1 XGBoost

Before building the model, all categorical predictor variables were converted into numerical forms using the One-Hot Encoding technique, except for the Spending Score variable, which was manually converted using Label Encoding. This was done to avoid the assumption of an ordinal relationship between categories that are nominal. Meanwhile, the target variable Segmentation was transformed into numerical form using Label Encoding because XGBoost cannot accept labels in text format for multiclass classification. Additionally, this is because the XGBoost model only accepts targets in numerical form, not as strings or booleans, as in One-Hot Encoding.

Once the data was prepared, hyperparameter tuning was performed using a Randomized Search Cross Validation approach to efficiently find the best parameter combination without having to evaluate all possible combinations. This method allows for a faster and more effective search compared to grid search, especially in large parameter spaces. The tuning results show the optimal parameter combination, shown in Table 5. These parameters collectively help regulate the model's complexity, reduce the risk of overfitting, and improve generalization ability by assigning optimal weights to each decision tree built.

Table 5. Optimal Parameter Combination of XGBoost Model Using All Variables

Parameter	Combination of Values	Optimal Value
subsample	0.6; 0.7; 0.8; 0.9; 1.0	0.7
scale_pos_weight	1, 5, 10, 20	5
reg_lambda	0; 0.1; 1.0; 10	0
reg_alpha	0; 0.1; 1.0; 10	0.1
n_estimators	100, 200, 300, 400, 500	200
max_depth	3, 5, 7, 9, 12	5
learning_rate	0.01; 0.05; 0.1; 0.2; 0.3	0.05
gamma	0; 0.1; 0.2; 0.5	0.5
colsample_bytree	0.4; 0.6; 0.8; 1.0	0.6
colsample_bylevel	0.4; 0.6; 0.8; 1.0	0.8

After the XGBoost model is constructed using the best parameters obtained from the tuning process, the next step is to evaluate its performance on both the training and testing datasets. This evaluation is conducted to assess how well the model has learned patterns from the training data and how effectively it can generalize and make predictions on unseen data. With an ROC AUC of 0.8435, it shows that the XGBoost model built using all variables has a very good ability to distinguish between customer segment classes. The closer the value is to 1, the better the model performs in classification tasks. A score above 0.80, as shown in this result, suggests that the model successfully captures key patterns in the training data and classifies with a high level of accuracy.

After evaluating the model on the training data, the next step is to test the model's performance on the testing data to measure its ability to generalize to unseen data that was not used during the training process. The ROC AUC value of 0.5837 on the testing data shows that the model's ability to distinguish between customer segments is moderate but relatively low compared to the training performance. This suggests that the model's generalization to unseen data is limited, with a success rate of approximately 58.37% in correctly differentiating between segments. The decrease in ROC AUC from training to testing data may show some degree of overfitting, where the model fits the training data well but performs less effectively on new data.

3.3.2 LightGBM

Before building the LightGBM model, all predictor variables were first converted into numerical form. For categorical variables, the One-Hot Encoding technique was used, with some exceptions manually mapped using Label Encoding to preserve the data structure. This approach ensures that the model can effectively utilize information from all variables without assuming inappropriate ordinal relationships.

Once the data was prepared, hyperparameter tuning was performed using a Randomized Search Cross Validation approach to efficiently find the best parameter combination without evaluating all possible combinations. This method allows for faster search with competitive results, especially in large parameter spaces. The tuning results show the optimal parameter combination for LightGBM as presented in Table 6. These parameters help regulate the model's complexity, reduce the risk of overfitting, and improve generalization ability by assigning optimal weights to each decision tree constructed.

Table 6. Optimal Parameter Combination of LightGBM Model Using All Variables

Parameter	Combination of Values	Optimal Value
subsample	0.6; 0.7; 0.8; 0.9; 1.0	0.8
reg_lambda	0; 0.1; 1.0; 10	10
reg_alpha	0; 0.1; 1.0; 10	0
num_leaves	15, 31, 63, 127, 255	15

Parameter	Combination of Values	Optimal Value
n_estimators	100, 200, 300, 400, 500	500
min_child_samples	10, 20, 30, 50	50
max_depth	-1, 5, 10, 15, 20, 30	30
learning_rate	0.01; 0.05; 0.1; 0.2; 0.3	0.01
colsample_bytree	0.4; 0.6; 0.8; 1.0	0.6
class_weight	*None, balanced*	*balanced*

After building the LightGBM model with the best-tuned parameters, with an ROC AUC value of 0.8202 on the training data, it indicates that the LightGBM model has a very good ability to distinguish between customer segment classes. This value means the model can correctly differentiate approximately 82.02% of randomly selected pairs of customers from different segments. The closer the value is to 1, the better the model performs in classification. A score above 0.8 suggests that the model effectively captures important patterns in the training data and classifies with high accuracy.

After evaluating the model on the training data, the evaluation on the testing data results in a value of 0.5833. It shows that the LightGBM model's ability to distinguish between customer segment classes is moderate to low. This suggests that the model has limitations in generalizing new, unseen data during training. The value implies a success rate of approximately 58.33% in differentiating between customer segments. The performance decline compared to the training data may show overfitting.

3.3.3 CatBoost

In the modeling process using CatBoost, all predictor variables, including categorical ones, can be used directly without requiring additional encoding steps such as One-Hot Encoding or Label Encoding. This advantage of CatBoost lies in its ability to automatically handle categorical variables, which simplifies the pre-modeling process and minimizes the risk of misinterpreting relationships among categories. Additionally, the target variable "Segmentation" undergoes Label Encoding to convert it into a numerical format suitable for multiclass classification. Although CatBoost supports categorical processing, labeling the target numerically ensures format consistency during the evaluation process.

Once all data is prepared, hyperparameter tuning is conducted to find the optimal model parameters. This tuning process uses a Randomized Search Cross Validation approach, allowing efficient searching for the best parameter combination without exploring the entire parameter space (only exploring parameters within a predefined search space. Based on the tuning results, the optimal parameter combination is presented in Table 7.

Table 7. Optimal Parameter Combination of CatBoost Model Using All Variables

Parameter	Combination of Values	Optimal Value
iterations	100, 200, 300, 400, 500	300
depth	4, 6, 8, 10, 12	6
learning_rate	0.01; 0.05; 0.1; 0.2; 0.3	0.1
12_leaf_reg	1, 3, 5, 7, 9	7
bagging_temperature	0; 0.5; 1; 2	0
colsample_bylevel	0.4; 0.6; 0.8; 1.0	0.4
border_count	32, 64, 128, 254	32
random_strength	0; 0.5; 1; 2	0.5

The CatBoost model is then retrained using the best-tuned parameters and evaluated on both training and testing data to measure the model's ability to learn patterns and generalize to new data. The AUC value of 0.8258 on the training data indicates that the CatBoost model built using all variables has a very good ability to distinguish between customer segment classes on the training data. This value

means that the model can accurately differentiate approximately 82.58% of randomly selected pairs of customers from different segments. With an ROC AUC close to 1, the model demonstrates high classification performance and strong learning capability on the training data.

The next evaluation was conducted on the testing data to figure out how well the model can classify new, unseen data. With an ROC AUC of 0.5759 on the testing data, it shows that the CatBoost model's ability to distinguish between customer segment classes on new data is relatively low to moderate. This suggests that the model has limitations in generalizing patterns from the training data to the testing data, with an approximate success rate of 57.59% in differentiating between customer segments. This decline in performance may indicate overfitting or that the model has yet to capture sufficient variation in the testing data.

3.4 Modeling of Significant Variables

3.4.1 XGBoost

Based on the Likelihood Ratio Test for numerical variables and the Chi-square test for categorical variables at a 5% significance level conducted previously, the variables found to have a significant influence on the Segmentation variable include: Age, Family Size, Gender, Ever Married, Graduated, Profession, Spending Score, and Var 1. Subsequently, modeling was performed using the XGBoost algorithm, utilizing only these significant variables.

After obtaining the optimal parameter combination, the next step is to build a model using the XGBoost algorithm with those parameters. The model is trained using the training data, followed by separate evaluations of its performance on the training and testing datasets. With ROC AUC value 0.8283, it shows that the XGBoost model built using the significant variables has a very good ability to distinguish between customer segment classes. This value means that the model can correctly differentiate between two randomly chosen customers from different segments with an approximate success rate of 82.83%. An ROC AUC value closer to 1 shows better model performance in classification. Values above 0.8, as shown in this result, show that the model is able to recognize important patterns in the training data and classify the data with a fairly high level of accuracy.

Subsequently, an evaluation was conducted on the testing data to assess the extent to which the model is capable of generalizing to new, unseen data. The ROC AUC value on the testing data is 0.5817, which indicates that the model has low classification capability when applied to new, unseen data. This value is only slightly higher than a random guess (0.5), suggesting that the model's performance has significantly declined compared to when it was evaluated on the training data. This means that the model is less able to consistently distinguish between customers from different segments in unfamiliar data, resulting in unstable prediction accuracy. The noticeable drop in performance also suggests a potential overfitting issue, where the model may have adapted too closely to the training data and failed to capture general patterns that apply beyond it.

3.4.2 LightGBM

Based on the initial analysis results using the Likelihood Ratio Test for numerical variables and the Chi-Square Test for categorical variables at a significance level of 5%, eight variables were found to be significant for the target variable Segmentation, namely Age, Family Size, Gender, Ever Married, Graduated, Profession, Spending Score, and Var 1. Next, the Light Gradient Boosting Machine (LightGBM) algorithm was used, involving only these significant variables as predictors.

After obtaining the optimal parameter combination, the next step is to perform modeling using the LightGBM algorithm with those parameters. The model is trained using training data, and then the model's performance is evaluated using training and testing data separately. The ROC AUC value shows that the model has good classification capabilities, with an accuracy rate of 81.72% in distinguishing between positive and negative classes. This indicates that the model is significantly superior to random predictions but has not yet reached the excellent category. To ensure the reliability of the model, further validation of the test data is necessary to avoid overfitting. If significant differences are found between the performance of the training data and the test data, it is recommended to perform

hyperparameter optimization or feature improvement. Overall, this model is reliable for practical applications, but other evaluations are needed to ensure its performance consistency.

Next, an evaluation was conducted on the testing data to measure how well the model can generalize to new data that has never been seen before. The ROC AUC value of 0.5829 indicates that the model has very limited discriminatory ability in distinguishing between positive and negative classes. With a value only slightly higher than the random baseline (0.5), this model does not provide significantly better predictions than random guessing. These results suggest several possibilities, including that the features used do not have a strong relationship with the target variable, the selected model is not suitable for the data characteristics, or better data preprocessing is required. Models with an AUC value below 0.7 are generally considered inadequate, so it is recommended to re-evaluate feature selection, model algorithms, and feature engineering processes before proceeding to further validation stages.

3.4.3 CatBoost

Based on the Likelihood Ratio Test for numerical variables and the Chi-square test for categorical variables at a 5% significance level conducted previously, it was found that the variables significantly affecting the Segmentation variable are: Age, Family Size, Gender, Ever Married, Graduated, Profession, Spending Score, and Var 1. Subsequently, modeling was performed using the CatBoost algorithm by including only these significant variables.

After obtaining the optimal parameter combination, modeling was carried out using the CatBoost algorithm with these parameters. The model was trained using the training data, followed by separate evaluations on both the training and testing datasets. The ROC AUC value of 0.8293 indicates that the CatBoost model has strong classification performance on the training data. This value means that there is an 82.93% probability that the model can correctly distinguish between randomly selected instances from different classes. Since ROC AUC values range from 0.5 (no discrimination) to 1.0 (perfect discrimination), a score above 0.8 is generally considered to reflect good model performance. In this case, the model demonstrates a solid ability to separate customer segments based on the available features, showing consistent and reliable behavior when trained on the given data.

Next, an evaluation was carried out on the testing data to assess the model's ability to generalize to unseen data. The ROC AUC value of 0.5773 in the testing data indicates that the model has a relatively weak classification performance when applied to new, unseen data. This value is only slightly higher than 0.5, which represents random guessing, suggesting that the model has a limited ability to distinguish between different customer segments outside of the training data. A score in this range reflects that the model's predictive power is low in real-world scenarios, and its decision boundaries may not generalize well beyond the training set. The noticeable gap between training and testing performance also suggests a potential overfitting issue, where the model may have learned the training data too closely and failed to capture general patterns that apply to new data.

3.5 Comparison of All Models

The final evaluation was carried out by comparing the performance of all models based on the ROC AUC values obtained on the testing data, for both models that used all predictor variables (all variables) and those that only included significant variables. The comparison is shown in Table 8. The ROC AUC was chosen as the main evaluation metric because it provides an objective overview of the model's ability to distinguish between classes, even when the data is imbalanced, and it is not influenced by a specific classification threshold.

Table 8. Comparison of ROC AUC Scores on Testing Set

Models	ROC AUC on Testing Set	
	All Variables	Significant Variables
XGBoost	0.5837	0.5817
LightGBM	0.5834	0.5829
CatBoost	0.5759	0.5773

In general, all models demonstrated the best classification capability for class D but struggled to differentiate between classes A, B, and C, which share overlapping features. All three models showed relatively similar performance with a tendency toward overfitting, as indicated by the large gap between the training and testing results. However, based on the overall evaluation, it can be concluded that the XGBoost model using all available variables is the best model in this study, achieving the highest ROC AUC score of 0.5837 on the testing data.

Although the ROC AUC score on the testing data is still relatively low and only slightly above the random guess benchmark (0.5), this model was still selected as the best because it demonstrated the most stable performance among all alternatives tested. The low score may be attributed to the high similarity of characteristics among the segments, the complexity of the data, and the limited information provided by the available features, making it difficult for the model to consistently separate the classes in new, unseen data. Therefore, this XGBoost model is considered the most optimal option for this study, although further development is still needed, such as adding new variables or improving feature quality, to significantly enhance classification performance.

3.6 Best Model

Feature Importance was conducted to assess the relative contribution of each input variable within the XGBoost Model. The results, as illustrated in Fig. 2, indicate that the Profession variable holds the highest importance. This indicates that someone's profession serves as a highly informative feature for predicting customer segmentation outcomes, as it encapsulates various socio-economic factors such as income level, job stability, daily mobility demands, and lifestyle patterns, which directly influence automotive preferences and purchasing behavior.

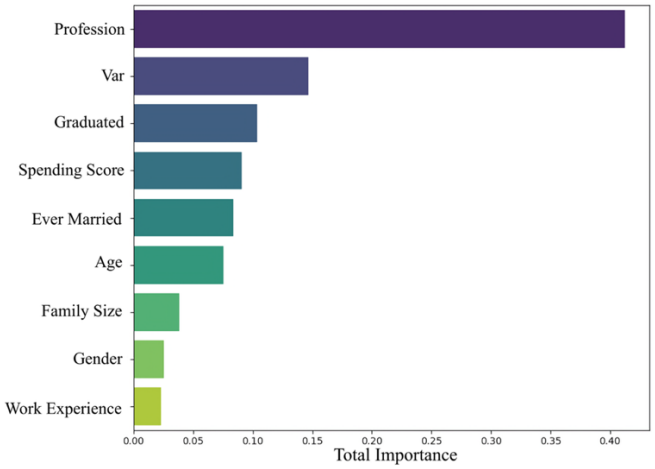

Figure 2. Feature Importance of Best Model

From the best model results using XGBoost with all variables, the characteristics of each customer segmentation class can be seen in Table 9. Based on the analysis results, customers in this segment A are on average around 45-46 years old, have around 2 years and 10 months of work experience, and live in small families with an average of 2 family members, including themselves. This shows that most customers in this segment are of a productive age but are still relatively new to the world of work. In terms of categorical characteristics, most customers are male, married, and have a college education. In terms of occupation, the most common profession in this segment is artist. This profession indicates that customers in this segment tend to have a creative background, with different mindsets, lifestyles, and consumption preferences from other professions.

Table 9. Class Segmentation Characteristics on Best Model

Variable	Class			
	A	B	C	D
Mean Age	45.857	51.047	53.626	30.115

Variable	Class			
	A	B	C	D
Mean Work Experience	2.885	1.924	1.880	2.609
Mean Family Size	2.003	2.603	2.903	3.492
Gender Mode	Male	Male	Male	Male
Ever Married Mode	Yes	Yes	Yes	No
Graduated Mode	Yes	Yes	Yes	No
Profession Mode	Artist	Artist	Artist	Healthcare
Spending Score Mode	Low	Low	Average	Low
Var 1 Mode	Cat_6	Cat_6	Cat_6	Cat_6

4. CONCLUSION

Based on the comprehensive analysis conducted in this study, it can be concluded that the application of advanced gradient boosting algorithms, specifically XGBoost, LightGBM, and CatBoost, enabled the effective classification of potential new customers into four distinct segments, each with unique demographic and behavioral characteristics. Among the evaluated models, XGBoost demonstrated superior performance as measured by the ROC AUC metric, which is particularly robust in the presence of class imbalance, and thus was selected as the best model for customer segmentation. Feature importance analysis revealed that variables such as Profession, Var 1, and Graduated played the most significant roles in segment differentiation, while factors like Spending Score and Age also contributed meaningfully, and variables such as Ever Married, Gender, Work Experience, and Family Size had comparatively lower influence. The resulting customer segments, ranging from productive-age, artistically inclined, and frugal individuals (Segment A), to mature customers with selective preferences (Segment B), highly educated seniors with moderate consumption (Segment C), and young, family-oriented, cautious spenders (Segment D), offer actionable insights for targeted marketing. It is therefore recommended that the company adopt the XGBoost-based segmentation model for market expansion initiatives, using the identified key variables to craft personalized marketing strategies tailored to the specific needs and preferences of each segment. This data-driven, personalized approach is expected to enhance marketing effectiveness, foster stronger customer relationships, and support the company's strategic objectives in new markets by ensuring that outreach efforts are both precise and impactful.

Author Contributions

Novri Suhermi: Conceptualization, Methodology, Supervision, and Writing–Review and Editing. Rahida Rihhadatul Aisy: Data Curation, Formal Analysis, and Writing– Original Draft. Aulia Afifatur Rohmah: Software Implementation, Validation, Visualization. Anis Alif Nurhayati: Data Preprocessing, Literature Review. Agnes Nathania Pramesty: Statistical Analysis, Interpretation of Results. Aura Lovi Ardanika: Model Training, Hyperparameter Tuning. Fauziyah Nurul Isnaini: Performance Evaluation, Proofreading. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research was supported by Institut Teknologi Sepuluh Nopember (ITS) under the Department Research Funding scheme.

Acknowledgment

The authors would like to thank the reviewers for their valuable comments and suggestions, which have greatly enhanced the quality of this paper. The authors also acknowledge the financial support from Institut Teknologi Sepuluh Nopember under the Department Research Funding scheme.

Declarations

The authors declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Generative AI and AI-assisted Technologies

Generative AI tools (e.g., ChatGPT) were used solely for language refinement, including grammar, spelling, and clarity. The scientific content, analysis, interpretation, and conclusions were developed entirely by the authors. All final text was reviewed and approved by the authors.

REFERENCES

[1] Y. Afrida, "ANALISIS PEMBIAYAAN MURABAHAH DI PERBANKAN SYARIAH," *Jurnal Ekonomi dan Bisnis Islam (JEBI)*, vol. 1, no. 2, pp. 155–166, 2016.

[2] K. Suresh, "CUSTOMER SEGMENTATION CLASSIFICATION," https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation.

[3] W. Nugraha and M. Syarif, "TEKNIK WEIGHTING UNTUK MENGATASI KETIDAKSEIMBANGAN KELAS PADA PREDIKSI CHURN MENGGUNAKAN XGBOOST, LIGHTGBM, DAN CATBOOST," *Techno.Com*, vol. 22, no. 1, pp. 97–108, Feb. 2023. doi: https://doi.org/10.33633/tc.v22i1.7191

[4] S. E. Herni Yulianti, Oni Soesanto, and Yuana Sukmawaty, "PENERAPAN METODE EXTREME GRADIENT BOOSTING (XGBOOST) PADA KLASIFIKASI NASABAH KARTU KREDIT," *Journal of Mathematics: Theory and Applications*, pp. 21–26, Aug. 2022. doi: https://doi.org/10.31605/jomta.v4i1.1792

[5] I. Z. A. Illah, W. S. J. Sapu, and A. T. Damaliana, "IMPLEMENTASI METODE KLASIFIKASI LIGHTGBM DAN ANALISIS SURVIVAL DALAM MEMPREDIKSI PELANGGAN CHURN," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 8, no. 1, pp. 43–53, Jun. 2024. doi: https://doi.org/10.31603/komtika.v8i1.11194

[6] J. Lin, "APPLICATION OF MACHINE LEARNING IN PREDICTING CONSUMER BEHAVIOR AND PRECISION MARKETING," *PLoS One*, vol. 20, no. 5, p. e0321854, May 2025. doi: https://doi.org/10.1371/journal.pone.0321854

[7] A. Geron, *HANDS-ON MACHINE LEARNING WITH SCIKIT-LEARN, KERAS, AND TENSORFLOW, 2nd Edition*. O'Reilly Media, Inc., 2019.

[8] Z. Mustaffa and M. H. Sulaiman, "ADVANCED FORECASTING OF BUILDING ENERGY LOADS WITH XGBOOST AND METAHEURISTIC ALGORITHMS INTEGRATION," *Energy Storage and Saving*, Aug. 2025. doi: https://doi.org/10.1016/j.enss.2025.03.005

[9] A. R. Zaidi, T. Abbas, A. Daud, O. Alghushairy, H. Dawood, and N. Sarwar, "ENHANCING ANDROID MALWARE DETECTION WITH XGBOOST AND CONVOLUTIONAL NEURAL NETWORKS," *Computers, Materials & Continua*, vol. 84, no. 2, pp. 3281–3304, 2025. doi: https://doi.org/10.32604/cmc.2025.063646

[10] X. Song, J. Shi, C. Zhu, F. Xian, Z. Dong, and J. Li, "XGBOOST MACHINE LEARNING ALGORITHM FOR PREDICTING UNPLANNED READMISSION IN ELDERLY PATIENTS WITH CORONARY HEART DISEASE," *Geriatr Nurs (Minneap)*, vol. 66, p. 103609, Nov. 2025. doi: https://doi.org/10.1016/j.gerinurse.2025.103609

[11] N. Qin, *et al.*, "FORECASTING THE MECHANICAL COMPACTION INFLUENCE ON SOYBEAN YIELD USING XGBOOST-ANN," *Information Processing in Agriculture*, Sep. 2025. doi: https://doi.org/10.1016/j.inpa.2025.09.002

[12] S.-K. Di, Y.-Y. Wang, D. Yang, Y.-H. Liu, J. Zhang, and W.-Z. Zheng, "SMOTE-ENHANCED XGBOOST FOR RAPID SEISMIC DAMAGE ASSESSMENT OF BRIDGE PORTFOLIOS," *Soil Dynamics and Earthquake Engineering*, vol. 199, p. 109712, Dec. 2025. doi: https://doi.org/10.1016/j.soildyn.2025.109712

[13] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*, 3rd ed. Birmingham: Packt Publishing., 2019.

[14] L. Zhang, H. Liu, and Z. Fan, "GRADIENT BOOSTING MACHINES AND THEIR APPLICATIONS IN CLASSIFICATION TASKS: AN OVERVIEW," *ACM Comput. Surv.*, vol. 54, no. 5, 2021.

[15] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "DIAGNOSIS OF DIABETES MELLITUS USING GRADIENT BOOSTING MACHINE (LIGHTGBM)," *Diagnostics*, vol. 11, no. 9, p. 1714, Sep. 2021. doi: https://doi.org/10.3390/diagnostics11091714

[16] H. Talebi, A. K. Bardsiri, and V. K. Bardsiri, "DEVELOPING A HYBRID MACHINE LEARNING MODEL FOR EMPLOYEE TURNOVER PREDICTION: INTEGRATING LIGHTGBM AND GENETIC ALGORITHMS,"

Journal of Open Innovation: Technology, Market, and Complexity, vol. 11, no. 2, p. 100557, Jun. 2025. doi: https://doi.org/10.1016/j.joitmc.2025.100557

[17] R. Fatahi, H. Abdollahi, M. Noaparast, and M. Hadizadeh, "MODELING PROCESS CONTROL VARIABLES OF A CEMENT VERTICAL ROLLER MILL USING LIGHTGBM: FEED RATE AND MAIN DRIVE POWER," *Chemical Engineering Research and Design*, vol. 219, pp. 595–610, Jul. 2025. doi: https://doi.org/10.1016/j.cherd.2025.06.019

[18] Z. Zhang, *et al.*, "EXHAUST EMISSIONS PREDICTION IN SPARK IGNITION ENGINE USING LIGHTGBM OPTIMIZED WITH THE MARINE PREDATORS ALGORITHM," *Appl Therm Eng*, vol. 275, p. 126800, Sep. 2025. doi: https://doi.org/10.1016/j.applthermaleng.2025.126800

[19] A. Ampountolas and S. AlGharbi, "AN INNOVATIVE HYBRID LIGHTGBM-BPNN MODEL FOR ENHANCED COMMODITY FORECASTING ACCURACY," *Finance Research Open*, vol. 1, no. 1, p. 100004, Mar. 2025. doi: https://doi.org/10.1016/j.finr.2025.100004

[20] M. Li, H. Tao, M. Liu, and T. He, "STUDY ON ENHANCED FAULT DIAGNOSIS OF CHILLER UNITS IN HVAC SYSTEMS UNDER THE IMBALANCED DATA ENVIRONMENT USING GA-OPTIMIZED LIGHTGBM," *Energy Build*, vol. 330, p. 115360, Mar. 2025. doi: https://doi.org/10.1016/j.enbuild.2025.115360

[21] Y. Zhang, C. Zhu, and Q. Wang, "LIGHTGBM-BASED MODEL FOR METRO PASSENGER VOLUME FORECASTING," *IET Intelligent Transport Systems*, vol. 14, no. 13, pp. 1815–1823, Dec. 2020. doi: https://doi.org/10.1049/iet-its.2020.0396

[22] J. T. Hancock and T. M. Khoshgoftaar, "CATBOOST FOR BIG DATA: AN INTERDISCIPLINARY REVIEW," *J Big Data*, vol. 7, no. 1, p. 94, Dec. 2020. doi: https://doi.org/10.1186/s40537-020-00369-8

[23] M. T. Syamkalla, S. Khomsah, and Y. S. R. Nur, "IMPLEMENTASI ALGORITMA CATBOOST DAN SHAPLEY ADDITIVE EXPLANATIONS (SHAP) DALAM MEMPREDIKSI POPULARITAS GAME INDIE PADA PLATFORM STEAM," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 4, pp. 777–786, Aug. 2024. doi: https://doi.org/10.25126/jtiik.1148503

[24] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CATBOOST: UNBIASED BOOSTING WITH CATEGORICAL FEATURES," in *Neural Information Processing Systems*, 2018, pp. 6639–6649.

[25] M. H. Sulaiman, Z. Mustaffa, A. S. Samsudin, A. I. Mohamed, and M. M. Saari, "ELECTRIC VEHICLE BATTERY STATE OF CHARGE ESTIMATION USING METAHEURISTIC-OPTIMIZED CATBOOST ALGORITHMS," *Franklin Open*, vol. 11, p. 100293, Jun. 2025. doi: https://doi.org/10.1016/j.fraope.2025.100293

[26] I. Rehan, M. U. Rehman, M. Aamir, and S. Islam, "A CATBOOST AND EXTRATREES-BASED SOFTVOTING ENSEMBLE APPROACH FOR NON-INVASIVE DIABETES DETECTION USING HAIR LIBS SPECTRAL DATA," *Microchemical Journal*, vol. 217, p. 114980, Oct. 2025. doi: https://doi.org/10.1016/j.microc.2025.114980

[27] M. Shehab, R. Taherdangkoo, and C. Butscher, "A PHYSICS-BASED CATBOOST MODEL FOR WATER RETENTION OF COMPACTED BENTONITE WITH GLOBAL SENSITIVITY ANALYSIS," *Appl Clay Sci*, vol. 277, p. 107948, Dec. 2025. doi: https://doi.org/10.1016/j.clay.2025.107948

[28] H. Lavaei, M. Esmaeili, and M. Mehraein, "ENHANCED PREDICTION OF SCOUR DIMENSIONS: TEMPORAL VARIATIONS INDUCED BY TURBULENT PLANE WALL JETS USING FFNN, CATBOOST, AND XGBOOST MODELS," *Ocean Engineering*, vol. 333, p. 121539, Jul. 2025. doi: https://doi.org/10.1016/j.oceaneng.2025.121539

[29] B. So and E. A. Valdez, "ZERO-INFLATED TWEEDIE BOOSTED TREES WITH CATBOOST FOR INSURANCE LOSS ANALYTICS," *Appl Soft Comput*, vol. 177, p. 113226, Jun. 2025. doi: https://doi.org/10.1016/j.asoc.2025.113226

[30] T. Fawcett, "AN INTRODUCTION TO ROC ANALYSIS," *Pattern Recognit Lett*, vol. 27, no. 8, pp. 861–874, Jun. 2006. doi: https://doi.org/10.1016/j.patrec.2005.10.010

[31] D. M. W. Powers, "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS AND CORRELATION," *International Journal of Machine Learning Technology* , vol. 2, no. 1, 2011.