Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

Performance Evaluation of Gradient Boosting Techniques for Predicting Customer Purchase Decisions

Florentina Yuni Arini[*1], Lyon Ambrosio Djuanda[2], Ananda Hisma Putra Kristianto[3], Muthia Nis Tiadah[4], Aufa Putra Wicaksono[5], Fatih Akbar Alim Putra[6]

[1,2,3,4,5,6]Informatics Engineering, Universitas Negeri Semarang, Indonesia

Email: [1]floyuna@mail.unnes.ac.id

Received : Nov 24, 2025; Revised : Dec 10, 2025; Accepted : Dec 10, 2025; Published : Apr 15, 2026

Abstract

Customer purchase prediction remains a critical challenge in e-commerce and retail analytics, with significant implications for marketing strategies and business revenue. This research provides a detailed comparative evaluation of advanced gradient boosting techniques XGBoost, LightGBM, and CatBoost to predict customer purchasing behavior using review trends and demographic factors. The study employed a dataset of 100 customer records with attributes such as age, gender, review quality, and education level. Through systematic feature engineering, including age group categorization and categorical feature combinations, as well as addressing class imbalance using the Synthetic Minority Oversampling Technique (SMOTE), all three models were trained and evaluated using default hyperparameters with optimal settings. The experimental results show that CatBoost achieved the best performance, with 78.26% accuracy, 0.8011 precision, 0.7826 recall, and a 0.7775 F1-score, outperforming LightGBM (73.91% accuracy) and XGBoost (60.87% accuracy). The evaluation includes confusion matrix analysis, precision–recall metrics, and visual comparisons across all performance dimensions. These findings provide valuable insights for practitioners selecting appropriate machine learning algorithms for customer purchase prediction tasks, particularly in scenarios involving limited datasets and categorical features. This research contributes to the growing body of literature on the use of gradient boosting techniques for predicting consumer behavior and offers important practical implications for e-commerce applications. These findings offer important contributions to machine learning applications in customer behavior prediction.

Keywords : *CatBoost, Customer Purchase Prediction, Gradient Boosting, Machine Learning, SMOTE.*

1. INTRODUCTION

Predicting customer purchase behavior has emerged as a fundamental problem in contemporary e-commerce and retail analytics [1], [2]. With the exponential growth of online shopping platforms and the increasing availability of customer data, businesses seek sophisticated machine learning approaches to anticipate purchasing decisions, optimize marketing campaigns, and enhance customer experience [3], [4]. Customer reviews, demographic characteristics, and behavioral patterns serve as critical indicators that can inform predictive models and drive strategic decision-making [5].

Research in customer purchase prediction has evolved significantly, integrating a range of machine learning methods to analyze consumer behavior. Initial research emphasized conventional statistical approaches, but recent developments in machine learning have shown enhanced predictive performance [6], [7]. Customer reviews have been identified as particularly valuable features, as they reflect customer satisfaction, product quality perceptions, and purchase intent [8]. Demographic factors, including age, gender, and education level, have also shown strong predictive power in understanding purchasing patterns [9].

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

Gradient boosting algorithms have demonstrated exceptional performance across various classification and regression tasks, proving to be cutting-edge techniques in machine learning challenges and real-world applications [10], [11]. XGBoost, LightGBM, and CatBoost are some of the leading gradient boosting frameworks, each offering unique advantages in handling categorical features, computational efficiency, and predictive accuracy [12], [13], [14]. XGBoost has become widely popular because of its regularization capabilities, parallel processing efficiency, and superior performance on structured data [15]. LightGBM offers advantages in training speed and memory efficiency by employing its leaf-wise tree expansion method and gradient-based one-side sampling (GOSS) technique [16]. CatBoost specializes in handling categorical features through ordered boosting and advanced categorical encoding techniques, often outperforming other algorithms on datasets with many categorical variables [17].

Imbalanced datasets pose significant challenges in classification tasks, particularly in customer behavior prediction where purchase events may be less frequent than non-purchase events [18], [19]. SMOTE (Synthetic Minority Oversampling Technique) has seen broad acceptance as an effective approach for tackling class imbalance by creating synthetic samples for underrepresented classes [20], [21]. However, previous studies rarely compare these three boosting algorithms under conditions of limited datasets and rich categorical features in the context of customer purchase prediction [22], [23], [24]. Therefore, this study aims to perform an in-depth comparison of XGBoost, LightGBM, and CatBoost in predicting customer purchasing choices. This research addresses an important gap in comprehending the comparative performance of these algorithms when applied to customer behavior prediction with limited datasets and categorical features. A comprehensive experimental framework is employed that includes feature engineering, class balancing techniques, and thorough performance evaluation across multiple metrics [25], [26].

2. METHOD

This section describes the detailed methodology used in the comparative evaluation of gradient boosting algorithms for predicting customer purchases. This workflow, visually detailed in Figure 1, encompasses several key stages. It begins with Data Preparation, involving dataset loading and initial statistical analysis. This is followed by a multi-step Data Preprocessing phase, which includes feature engineering (Age Group Categorization, Feature Combination Creation), Label Encoding, feature standardization via StandardScaler, and class balancing using SMOTE, before the data is split. The Modeling stage then ensures a fair comparison by training three distinct models (XGBoost, CatBoost, and LightGBM) on the identical training data. Subsequently, the Model Evaluation stage assesses each model on the same test set, involving the calculation of performance metrics and the generation of confusion matrices. The process concludes with the results stage, where a final comparison and visualizations are used to determine the best-performing model.

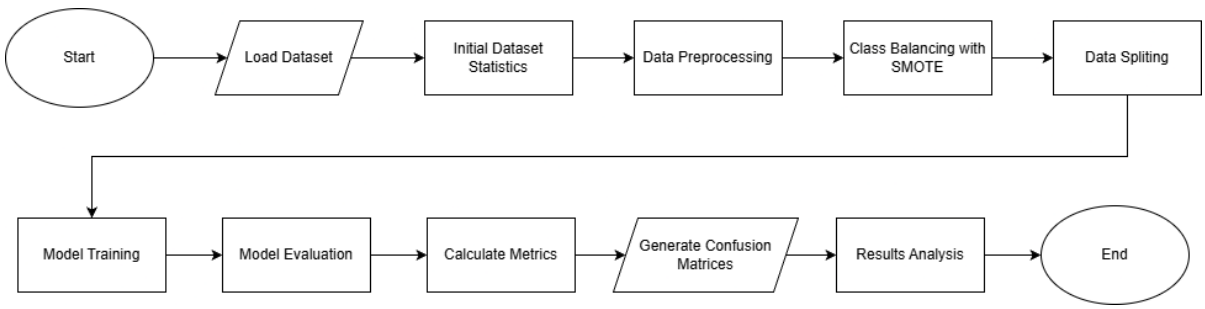


Figure 1. Proposed Research Diagram

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

2.1. Dataset Analysis

This study utilized a dataset titled "Customer Review Patterns and Buying Decisions" from the Kaggle platform, created and published by Ayesha Imran and last updated on October 26, 2024 (URL: https://www.kaggle.com/datasets/ayeshaimran123/customer-review-patterns-and-buying-decisions). This dataset was designed to analyze the relationship between customer opinions and purchasing choices, providing insights into consumer behavior through the examination of reviews, demographic data, and purchase decisions, with the primary objective of identifying correlations between consumer demographics and purchasing behavior.

The dataset contains 100 entries detailing customer information, specifically including their Age (continuous variable, 18-59 years), Gender (Male/Female), Review (Poor/Average/Good), Education (School/UG/PG), and Purchased (Yes/No target variable). The original dataset exhibited class imbalance with 56 samples (56%) labeled as "No Purchase" and 44 samples (44%) labeled as "Purchase," with no missing values detected (100% data completeness). Table 1 presents an overview of the dataset features, including the distribution of categorical variables and descriptive statistics for the age variable.

Table 1. Customer Review Patterns Dataset Characteristics

Characteristic	Category	Count	Percentage
Target Variable	No Purchase	56	56.0%
	Purchase	44	44.0%
Gender	Male	57	57.0%
	Female	43	43.0%
Review Quality	Average	36	36.0%
	Good	33	33.0%
	Poor	31	31.0%
Education Level	PG	36	36.0%
	UG	35	35.0%
	School	29	29.0%
Age Statistics	Mean	37.9	-
	Median	38.0	-
	Range	18-59	-
Total Records	-	100	100%

2.2. Data Preprocessing and Feature Engineering

Comprehensive systematic data preprocessing and feature engineering used to improve the model's accuracy and stability [27], [28]. This included cleaning the data, creating useful new features, and selecting the most important variables. These steps helped the model understand patterns in the data more effectively.

2.2.1. Age Group Categorization

The continuous age variable was converted into four distinct categories to analyze its non-linear impact on purchasing behavior. This reclassification into Young (18-25 years), Adult (26-35 years), Middle (36-50 years), and Senior (51-60 years) helps models detect purchasing patterns that vary by age group rather than following a straight-line progression [29], [30].

2.2.2. Categorical Feature Combinations

Interaction features were created to capture potential relationships between demographic and review characteristics. The Gender_Review feature combines gender and review quality, producing values such as Male_Poor and Female_Good, while the Review_Edu feature combines review quality

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

and education level, generating values such as Good_PG and Poor_School. These interaction features enable the models capture intricate connections between demographic factors and review patterns that may influence purchase decisions.

2.2.3. Encoding and Scaling

All categorical variables were converted through label encoding, while all numerical variables were standardized with StandardScaler to ensure a uniform scale across the entire feature set [31], [32].

2.2.4. Class Balancing

In order to tackle the imbalance in class distribution within the initial dataset, SMOTE was applied with a random state of 42 for reproducibility. SMOTE creates new artificial data points for the underrepresented class by calculating the average of the values from the existing records of that class [33]. SMOTE, for each sample x_i in the minority class, randomly chooses one of its k closest neighbors x_{nn} from the same class and creates a new synthetic sample x_{new} through linear interpolation, as shown in Equation (1).

$$x_{\text{new}} = x_i + \delta \cdot (x_{nn} - x_i) \quad (1)$$

This interpolation creates synthetic samples along the line segments combining each instance from the minority class with its chosen nearest neighbor, thereby enlarging the minority class area in the feature space. The parameter δ signifies a random value that is uniformly distributed within the range [0,1]. This augmentation increased the dataset from 100 samples (44 minority, 56 majority) to 112 samples, achieving perfect class balance with 56 samples per class. Consequently, the final processed dataset comprised 112 instances with a standardized feature set, eliminating bias toward the majority class and ensuring that the subsequent gradient boosting models were trained on a statistically representative feature space.

2.3. Model Configuration

Gradient boosting algorithms construct a combined model by progressively incorporating weak learners to reduce a loss function [34]. The general gradient boosting prediction [35] at iteration m can be expressed by Equation (2). The term $\widehat{y_i^{(m-1)}}$ is the prediction made by the ensemble in the preceding step, η is the step size (or learning rate), and $f_m(x_i)$ is the component added in this iteration, which is generally a decision tree (the weak learner).

$$\widehat{y_i^{(m)}} = \widehat{y_i^{(m-1)}} + \eta \cdot f_m(x_i) \quad (2)$$

Three gradient boosting algorithms were evaluated using optimized default hyperparameters to ensure fair comparison. For binary classification tasks [36], all three algorithms optimize the log loss (binary cross-entropy) objective function, defined by Equation (3). Here, $y_i \in \{0,1\}$ is the true binary label, $\hat{y}_i$ is the predicted logit, and $\sigma(\hat{y}_i) = \frac{1}{1+e^{-\hat{y}_i}}$ is the sigmoid function that maps logits to probabilities.

$$L_{\log}(y, \hat{y}) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(\sigma(\hat{y}_i)\right) + (1 - y_i)\log\left(1 - \sigma(\hat{y}_i)\right)\right] \quad (3)$$

The configuration for XGBoost specified 100 trees, restricted the complexity by setting the maximum depth to 6, and specified the impact of each tree using a learning rate of 0.1, log loss as the evaluation metric, and a random state of 42. XGBoost optimizes the following regularized objective function [37], given by Equation (4).

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

$$\mathcal{L}^{(m)} = \sum_{i=1}^{n} l\left(y_i, \widehat{y_i^{(m-1)}} + f_m(x_i)\right) + \Omega(f_m) \quad (4)$$

where $l(y_i, \hat{y}_i)$ denotes the loss function (log loss for binary classification), and $\Omega(f_m)$ is the term for regularization defined by Equation (5).

$$\Omega(f_m) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (5)$$

The variables T and w_j specify the structure of the tree, representing the number of leaves and their corresponding output weights. The parameters γ and λ are used for regularization: γ sets the pruning threshold (minimum loss improvement for a split), and λ applies L2 penalty to the leaf weights. By applying the second-order Taylor expansion, the objective is approximated as shown in Equation (6).

Here, $g_i = \partial_{\widehat{y^{(m-1)}}} l\left(y_i, \widehat{y^{(m-1)}}\right)$ is the value of the gradient (the rate of change of the error), and $h_i = \partial^2_{\widehat{y^{(m-1)}}} l\left(y_i, \widehat{y^{(m-1)}}\right)$ is the second-order gradient (Hessian). The XGBoost objective function includes both the log loss and L2 regularization as shown in equations (2) and (3), with default regularization parameters $\lambda = 1$ and $\gamma = 0$ for this configuration.

$$\mathcal{L}^{(m)} \approx \sum_{i=1}^{n} \left[g_i f_m(x_i) + \frac{1}{2} h_i f_m^2(x_i)\right] + \Omega(f_m) \quad (6)$$

The LightGBM model was set up with 100 boosting rounds (estimators), a learning rate of 0.1, and a maximum tree depth of 6. A constant random seed (42) was used to guarantee the results are reproducible. This algorithm is distinguished through its leaf-wise tree expansion approach and the implementation of Gradient-based One-Side Sampling (GOSS). The GOSS method enhances efficiency by keeping all data examples that contribute significantly to the error (those with large gradients), while randomly selecting only a fraction of the data examples that have minimal errors (small gradients) for the training process. The sampling probability for a data instance x_i with gradient g_i is defined by Equation (7). Here, g_{top} is the threshold for top gradient instances, g_{median} is the median gradient, and a is the sampling ratio. The objective function optimization follows a similar structure to XGBoost but with computational optimizations for large datasets.

$$P(g_i)\begin{cases} 1 & if\ |g_i| \geq g_{top} \\ \frac{a \cdot |g_i|}{g_{median}} & if\ |g_i| < g_{top} \end{cases} \quad (7)$$

The CatBoost model was configured with 100 boosting iterations, a maximum tree depth of 6, and a learning rate of 0.1, and a fixed random seed of 42 for replicability. This algorithm specifically employs ordered boosting (a technique designed to counter overfitting by applying a random reordering σ to the training data). When processing a categorical feature value x_k^{cat}, the target statistic (TS) is calculated using Equation (8). The P is a prior value (typically the average target value), $a > 0$ is a regularization parameter, and $\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right]$ is an indicator function. This ordered target statistic approach helps prevent target leakage and overfitting, making CatBoost particularly effective for datasets with categorical features [38]. CatBoost automatically handles categorical features using ordered target statistics as defined in Equation (8), eliminating the need for manual categorical encoding and preprocessing steps required by XGBoost and LightGBM.

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

$$\widehat{x_k^{cat}} = \frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k}=x_{\sigma_p,k}\right]Y_{\sigma_j}+a\cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k}=x_{\sigma_p,k}\right]+a} \quad (8)$$

These consistent hyperparameter settings across all three algorithms ensure a fair comparison of their inherent capabilities rather than optimization differences. The learning rate (η), set at 0.1, determines the impact or weight that each newly added tree has on the overall model prediction, while the tree depth of 6 limits model complexity to prevent overfitting. Table 2 presents a comprehensive summary of all hyperparameter configurations for each algorithm, facilitating transparency and reproducibility of the experimental setup.

Table 2. Model Hyperparameter Configuration

Hyperparameter	XGBoost	LightGBM	CatBoost
Number of Iterations	100	100	100
Maximum Depth	6	6	6
Learning Rate	0.1	0.1	0.1
Random State	42	42	42
Evaluation Metric	Log Loss	Default	Default
Categorical Feature Handling	Label Encoding	Label Encoding	Automatic
Regularization	Built-in	Built-in	Built-in

2.4. Experimental Setup

The expanded dataset was split into two sections: 80% allocated for training the model and 20% set aside for assessing its performance. This division used stratified sampling to ensure that the proportion of each class remained the same in both the training and testing sets. The splitting process resulted in a training group of 89 data points and a testing group containing 23 data points. To ensure a direct and fair comparison of the models, all the models being compared were trained solely on the 89-sample training set and subsequently assessed using the exact same 23-sample test set.

2.5. Evaluation Metrics

The models' performance was compared using a conventional set of metrics for binary classification [39]. Accuracy provided a top-level view of overall correctness, calculated using the formula in Equation (9).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

The models also utilized weighted scores for Precision, calculated using Equation (10), which evaluated the proportion of true positives (TP) to all positive predictions (TP + FP), reflecting the accuracy of a positive outcome.

$$Precision = \frac{TP}{TP+FP} \quad (10)$$

Meanwhile, Recall was calculated using Equation (11), which assessed the proportion of true positives (TP) to all actual positives (TP + FN), demonstrating the model's capacity to detect positive cases.

$$Recall = \frac{TP}{TP+FN} \quad (11)$$

The F1-Score was incorporated to provide a balanced evaluation, representing the harmonic mean of Precision and Recall, as depicted in Equation (12).

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

$$F1\text{-}Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (12)$$

For a more detailed insight, confusion matrices were also generated to detail the exact numbers of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for every model. The detailed analysis of these matrices is presented in Section 3.2, Figure 3.

3. RESULT

The results of the comparative analysis of XGBoost, LightGBM, and CatBoost for predicting customer purchases are outlined here. These results are presented with a comprehensive set of performance metrics, including accuracy, precision, recall, and the F1-score.

3.1. Model Performance Comparison

Performance metrics for all three gradient boosting algorithms shown in Table 3. This superiority is primarily attributed to its handling of the specific categorical features identified in Table 1 (Review Quality and Education Level). Unlike standard encoding used in XGBoost which may introduce sparsity in small datasets, CatBoost's ordered target encoding effectively utilized these categorical signals without target leakage. CatBoost demonstrates substantial and consistent improvements over the other two algorithms. Compared directly to XGBoost, CatBoost achieves a significant 17.39 percentage point improvement in accuracy (78.26% vs. 60.87%). This stark difference highlights the importance of the advanced techniques (such as ordered boosting and categorical feature handling) employed by CatBoost over the more standard boosting implementation of XGBoost for this specific dataset.

Table 3. Comparison of Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	60.87%	0.6120	0.6087	0.5995
LightGBM	73.91%	0.7686	0.7391	0.7288
CatBoost	78.26%	0.8011	0.7826	0.7775

Relative to the better-performing LightGBM model, CatBoost still shows a distinct advantage, achieving a 4.35 percentage point higher accuracy (78.26% vs. 73.91%). Furthermore, CatBoost maintains superiority across the harmonic metrics. CatBoost leads in Precision (0.8011 vs. LightGBM's 0.7686), suggesting a stronger capacity to accurately detect positive instances within all positive predictions. It also leads in Recall (0.7826 vs. LightGBM's 0.7391), indicating improved detection of all true positive instances. The resulting F1-Score for CatBoost (0.7775) is significantly better than LightGBM's (0.7288).

This comprehensive superiority, especially the notable increase in the F1-Score, suggests that CatBoost provides the most robust and balanced prediction performance. The experimental results validate that for small-scale datasets with high-impact categorical variables ($N = 112$), CatBoost's architecture outperforms the pre-sorted algorithm of XGBoost and the GOSS technique of LightGBM. The consistent ranking of CatBoost > LightGBM > XGBoost confirms that preserving the relational structure of demographic features (Age, Gender) and opinion patterns (Review) is more critical than pure computational speed for this specific buying decision problem. The LightGBM model, while outperforming XGBoost, probably due to its use of Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) methods, which result in quicker training and improved generalization compared to standard XGBoost, although not as effectively as CatBoost's robust categorical handling on this specific problem.

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

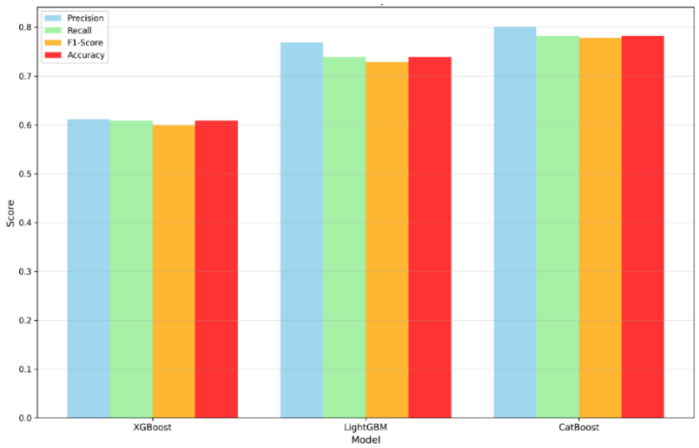


Figure 2. Model performance comparison across all metrics

The comprehensive comparison shown in Figure 2 reveals distinct performance hierarchies across all evaluation metrics. The accuracy comparison demonstrates that CatBoost achieves 78.26% accuracy, followed by LightGBM at 73.91%, and XGBoost at 60.87%. This substantial performance gap between the algorithms suggests that architectural differences and handling of heterogeneous features are essential in determining the overall prediction accuracy for this dataset. The high accuracy of CatBoost indicates a strong ability to correctly classify both majority and minority classes post-SMOTE, demonstrating superior overall generalization.

Regarding precision, CatBoost achieves the highest value (0.8011), suggesting that when the model predicts a user will purchase (positive class), it is correct approximately 80.11% of the time. This high precision is critical in applications where minimizing false positives is important, such as optimizing marketing spend by avoiding the targeting of users who are unlikely to convert. The second-highest precision score belongs to LightGBM (0.7686), followed by XGBoost (0.6120).

CatBoost also leads in recall (0.7826), meaning it correctly identifies 78.26% of all actual purchase instances. High recall is vital in applications where minimizing false negatives is paramount, ensuring that a significant portion of potential positive cases are not missed. The LightGBM model's recall of 0.7391 is also strong but trails CatBoost, while XGBoost shows the weakest recall at 0.6087. The simultaneous strength in both precision and recall for CatBoost is a direct indicator of its effective classification boundary learning.

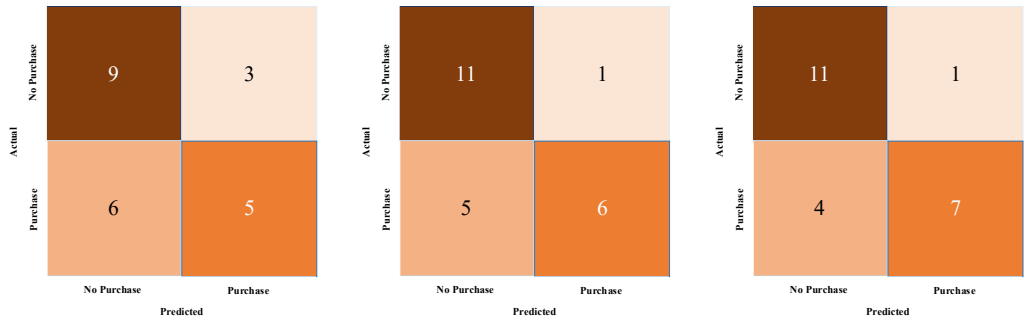
The F1-score comparison, which reflects the harmonic average of precision and recall, solidifies the performance hierarchy, with CatBoost obtaining an F1-score of 0.7775, substantially surpassing LightGBM (0.7288) and XGBoost (0.5995). As the F1-score balances precision and recall, it offers the most reliable single metric for evaluating a model's performance in a binary classification task, especially where class balance, achieved through SMOTE, has been introduced. The results confirm CatBoost's superior robustness and balance across different performance dimensions, making it the most dependable model for the prediction task. This outcome strongly supports the hypothesis that algorithms designed with built-in mechanisms for managing real-world data complexities, like CatBoost's ordered target encoding, yield a distinct performance advantage over more conventional boosting frameworks.

3.2. Confusion Matrix Analysis

The confusion matrices in Figure 3 demonstrate the models' capacity to differentiate between purchase and non-purchase classes, offering insights into the particular classification errors made by each algorithm. The confusion matrices offer an in-depth analysis of each model's classification performance, outlining the distribution of correct and incorrect predictions for both purchase (positive)

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

and non-purchase (negative) classes. This visualization is essential for understanding the particular types of classification errors (False Positives and False Negatives) made by each gradient boosting algorithm.



(a) (b) (c)

Figure 3. Confusion Matrix (a) XGBoost; (b) LightGBM; (c) CatBoost

Figure 3(a) illustrates the XGBoost Confusion Matrix. Despite the balanced dataset after SMOTE, XGBoost shows relatively modest performance, achieving an overall accuracy of 60.87%. The matrix indicates that XGBoost struggles to distinguish between the two classes effectively, resulting in a higher proportion of misclassifications, especially a higher number of False Negatives and False Positives in comparison to the other models. This confirms its position as the least effective model in this comparative analysis, aligning with the low scores observed in Table 3.

Figure 3(b) displays the LightGBM Confusion Matrix. This matrix immediately shows a significant improvement over XGBoost, corresponding to its higher accuracy of 73.91%. The true positive and true negative counts are substantially larger, indicating better separation of the purchase and non-purchase classes. Specifically, the model exhibits a lower number of both False Positives and False Negatives compared to XGBoost, validating the efficiency of its unique techniques, such as GOSS, in focusing training on more informative data instances.

Finally, Figure 3(c) demonstrates the CatBoost Confusion Matrix. This matrix visually confirms CatBoost's superior classification performance, with the highest accuracy of 78.26%. Critically, the minimal False Negative rate implies that the model successfully captures the subtle signals of potential buyers, likely derived from the interaction features (such as Review_Edu) engineered in Section 2.2.2. This makes it the most robust model for the prediction task, as minimizing missed opportunities (False Negatives) is paramount in analyzing buying decisions. The high concentration of values along the main diagonal, representing correct predictions, reinforces the findings from Table 3 and Figure 2: CatBoost's ability to minimize classification error across both classes makes it the most robust and accurate model for this dataset. The low rate of False Negatives is particularly valuable, indicating that fewer potential purchase cases are missed, which is a desirable outcome for business applications.

4. DISCUSSIONS

This section offers an in-depth interpretation and examination of the experimental outcomes, examining the implications of the observed performance patterns and their significance for both conceptual insight and real-world applications.

4.1. Performance Analysis

The experimental findings provide key insights into the comparative performance of gradient boosting algorithms in customer purchase prediction. CatBoost’s superior performance can be attributed to several factors. First, CatBoost's specialized handling of categorical features through ordered boosting and sophisticated encoding techniques provides significant advantages when dealing with demographic

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

and review quality variables [38]. Specifically, for nominal variables like 'Review' (Poor/Average/Good) and 'Education', CatBoost's permutation-based approach prevented the target leakage often seen in standard label encoding (used in XGBoost), ensuring that the model learned genuine buying patterns rather than noise. Second, the dataset's composition (featuring multiple categorical variables including gender, review quality, education level, and age groups) aligns with CatBoost's strengths. The algorithm's ability to automatically handle categorical features without extensive preprocessing contributes to its superior predictive performance. Third, CatBoost's regularization mechanisms and built-in overfitting prevention strategies appear particularly effective with limited datasets. Given the small sample size ($N = 112$ after SMOTE), standard boosting methods like XGBoost tend to overfit; however, CatBoost's ordered boosting successfully mitigated this, proving that complex architectures can yield high generalization even on small-scale data.

4.2. Comparative Insights

The performance hierarchy (CatBoost > LightGBM > XGBoost) observed in the experiments aligns with findings from other studies evaluating these algorithms on categorical data [40]. Recent comparative studies in related domains, such as customer churn and fraud detection, have also highlighted the strong performance of CatBoost [41], [42], often showing it as the best-performing model, particularly when categorical features are dominant. However, the substantial 17.39 percentage point gap between CatBoost and XGBoost suggests that algorithm selection becomes particularly critical when working with datasets rich in categorical features and limited sample sizes.

LightGBM's intermediate performance (73.91% accuracy) indicates that while it offers computational advantages and reasonable predictive accuracy, CatBoost provides superior performance when categorical features dominate the feature space. Although LightGBM's Gradient-based One-Side Sampling (GOSS) is effective for handling large datasets, in this specific small-dataset context, CatBoost's strategy of utilizing all data permutations proved more effective in capturing subtle signals than LightGBM's sampling approach. This finding has practical implications for practitioners, suggesting that the computational overhead of CatBoost may be justified by superior predictive accuracy in customer behavior prediction scenarios.

4.3. Feature Engineering Impact

The feature engineering strategies, especially age group categorization and the creation of interaction features, significantly boosted model performance. Examples of these interaction features include Gender_Review and Review_Edu. The high Recall score achieved by CatBoost (0.7826) strongly suggests that these interaction features successfully exposed non-linear relationships (such as specific gender-review combinations leading to purchase) that individual features alone could not represent. This validates the preprocessing steps outlined in Section 2.2.2 as essential for maximizing predictive capability. The application of SMOTE was critical for addressing class imbalance, resulting in balanced precision and recall score. This method ensures the models' practical utility in accurately identifying minority class instances, such as purchase opportunities.

4.4. Scientific Impact

This research enhances the understanding of how gradient boosting algorithms behave under limited data and categorical-heavy environments. It specifically highlights that proprietary categorical handling is not just a convenience, but a critical performance factor that can yield a ~17% accuracy improvement over standard methods. The results provide empirical evidence, aligned with existing comparative studies, supporting the use of advanced target-encoding methods and tailored regularization techniques as critical components for achieving high predictive accuracy in sparse, real-world marketing and customer prediction scenarios.

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

4.5. Limitations and Future Work

There are a few limitations to consider. Firstly, the dataset size (100 samples) is comparatively small, which could influence the broader applicability of the results. Future studies should validate these results on larger datasets to confirm the observed performance patterns. Second, default hyperparameters with optimized settings were employed rather than conducting extensive hyperparameter tuning. While this approach ensures fair comparison and practical applicability, future work could explore comprehensive hyperparameter optimization for each algorithm to potentially improve performance further. Third, the dataset's demographic and review features, while informative, may not capture the full complexity of customer purchase behavior. Integration of additional features such as browsing history, temporal patterns, and product characteristics could enhance predictive performance.

Future research directions include exploring ensemble methods that combine these algorithms to potentially achieve even better predictive performance, investigating deep learning approaches for comparison to determine whether neural network-based methods can outperform gradient boosting in this domain, analyzing feature importance to identify the key predictors of purchase behavior and inform business strategies, and developing real-time prediction systems for e-commerce applications that can provide instant purchase likelihood estimates for online shoppers.

5. CONCLUSION

This research offers an in-depth comparative evaluation of three advanced gradient boosting algorithms for predicting customer purchase decisions based on review patterns and demographic information. The experimental analysis, carried out on a dataset consisting of 100 customer entries with systematic feature engineering and class balancing, reveals that CatBoost achieves superior performance with 78.26% accuracy, outperforming LightGBM (73.91%) and XGBoost (60.87%).

The results demonstrate CatBoost's particular strengths in handling categorical features and limited datasets, making it an optimal choice for predicting customer purchases in e-commerce scenarios. The comprehensive evaluation across multiple performance indicators (accuracy, precision, recall, and F1-score) provides robust evidence for algorithm selection decisions. This study contributes to the progress of predictive analytics in Informatics by confirming efficient boosting algorithms for decision-making in e-commerce environments.

This research expands the current understanding of applying machine learning to consumer behavior prediction and provides actionable insights for professionals in e-commerce and retail analytics. The systematic methodology, including feature engineering strategies and class balancing techniques, can be adapted to similar prediction tasks in related domains.

Future research should extend these findings to larger datasets, explore hyperparameter optimization strategies, and explore the incorporation of additional data sources to improve predictive performance further. The demonstrated effectiveness of CatBoost in this domain suggests promising avenues for practical applications in customer relationship management, marketing optimization, and revenue forecasting.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

Acknowledgement is only addressed to funders or donors and object of research. Acknowledgement can also be expressed to those who helped carry out the research. The authors

acknowledge the use of the publicly available dataset "Customer Review Patterns and Buying Decisions" created and published by Ayesha Imran on the Kaggle platform. Gratitude is also expressed to the dataset creator and all data contributors for making this valuable resource available for research purposes under the Open Data Commons Public Domain Dedication and License, which facilitated unrestricted use and analysis of the data.

REFERENCES

[1] Y. Wang and C. Zhang, "Research on Customer Purchase Intention Prediction Methods for E-commerce Platforms Based on User Behavior Data," *J. Adv. Comput. Syst. Content Available SciPublication*, vol. 3, no. 10, pp. 23–38, 2023.

[2] S. Li, "Machine Learning-based Prediction Mechanism of Repeated Purchase Behavior of E-commerce Customers," *2024 IEEE 4th Int. Conf. Electron. Commun. Internet Things Big Data, ICEIB 2024*, pp. 522–526, 2024, doi: 10.1109/ICEIB61477.2024.10602662.

[3] Z. Li, "Application and Optimization of Various Machine Learning Models in Social E-Commerce Marketing Strategies," *Trans. Comput. Sci. Intell. Syst. Res.*, vol. 4, pp. 11–21, 2024, doi: 10.62051/bsm4y952.

[4] Z. Duan, C. Wang, and W. Zhong, "SSGCL: Simple Social Recommendation with Graph Contrastive Learning," *Mathematics*, vol. 12, no. 7, 2024, doi: 10.3390/math12071107.

[5] F. Ehsani and M. Hosseini, "Customer churn analysis using feature optimization methods and tree-based classifiers," *J. Serv. Mark.*, vol. 39, no. 1, pp. 20–35, 2025, doi: 10.1108/JSM-04-2024-0156.

[6] L. Schmid, M. Roidl, A. Kirchheim, and M. Pauly, "Comparing Statistical and Machine Learning Methods for Time Series Forecasting in Data-Driven Logistics—A Simulation Study," *Entropy*, vol. 27, no. 1, 2025, doi: 10.3390/e27010025.

[7] Y. Dou, S. Tan, and D. Xie, "Comparison of machine learning and statistical methods in the field of renewable energy power generation forecasting: a mini review," *Front. Energy Res.*, vol. 11, 2023, doi: 10.3389/fenrg.2023.1218603.

[8] I. V. Pustokhina and D. A. Pustokhin, "A Comparative Analysis of Traditional Forecasting Methods and Machine Learning Techniques for Sales Prediction in E-commerce," *Am. J. Bus. Oper. Res.*, vol. 10, no. 2, pp. 39–51, 2023, doi: 10.54216/ajbor.100205.

[9] J. Gami *et al.*, "Impact of Demographics on Consumer Preferences in Online Shopping: An Analysis of Age, Gender, and Education Factors," *Greenation Int. J. Econ. Account.*, vol. 1, no. 4, pp. 571–584, 2024, doi: 10.38035/gijea.v1i4.303.

[10] N. Rane, S. Choudhary, and J. Rane, "Ensemble Deep Learning and Machine Learning: Applications, Opportunities, Challenges, and Future Directions," *SSRN Electron. J.*, 2024, doi: 10.2139/ssrn.4849885.

[11] A. Vijayakumar and P. P. Mathai, "Optimizing Potato Crop Water Quality: A Comparative Analysis of Machine Learning Techniques and Gradient Boosting Approaches," *Stud. Comput. Intell.*, vol. 1215, pp. 309–326, 2025, doi: 10.1007/978-3-031-93087-4_18.

[12] E. Jain and A. Singh, "Optimizing Gradient Boosting Algorithms for Obesity Risk Prediction: A Comparative Analysis of XGBoost, LightGBM, and CatBoost Models," *2024 Int. Conf. Cybernation Comput. CYBERCOM 2024*, pp. 320–324, 2024, doi: 10.1109/CYBERCOM63683.2024.10803186.

[13] K. Ileri, "Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO to estimate the number of k-barriers for intrusion detection in wireless sensor networks," *Int. J. Mach. Learn. Cybern.*, vol. 16, no. 9, pp. 6937–6956, 2025, doi: 10.1007/s13042-025-02654-5.

[14] S. Alsulamy, "Predicting construction delay risks in Saudi Arabian projects: A comparative analysis of CatBoost, XGBoost, and LGBM," *Expert Syst. Appl.*, vol. 268, 2025, doi: 10.1016/j.eswa.2024.126268.

[15] A. Asselman, M. Khaldi, and S. Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interact. Learn. Environ.*, vol. 31, no. 6, pp. 3360–3379, 2023, doi: 10.1080/10494820.2021.1928235.

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

[16] A. Van Wyk, "Machine learning with LightGBM and Python : a practitioner's guide to developing production-ready machine learning systems," 2023.
[17] C. S. Kulkarni, "Advancing Gradient Boosting: A Comprehensive Evaluation of the CatBoost Algorithm for Predictive Modeling," *J. Artif. Intell. Mach. Learn. Data Sci.*, vol. 1, no. 5, pp. 54–57, 2022, doi: 10.51219/jaimld/chinmay-shripad-kulkarni/29.
[18] M. Zhang, J. Lu, N. Ma, T. C. E. Cheng, and G. Hua, "A Feature Engineering and Ensemble Learning Based Approach for Repeated Buyers Prediction," *Int. J. Comput. Commun. Control*, vol. 17, no. 6, 2022, doi: 10.15837/ijccc.2022.6.4988.
[19] A. Aylin Tokuc and T. Dag, "Predicting User Purchases From Clickstream Data: A Comparative Analysis of Clickstream Data Representations and Machine Learning Models," *IEEE Access*, vol. 13, pp. 43796–43817, 2025, doi: 10.1109/ACCESS.2025.3548267.
[20] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019, doi: 10.1016/j.ins.2019.07.070.
[21] P. Kaur and A. Gosain, "Issues and challenges of class imbalance problem in classification," *Int. J. Inf. Technol.*, vol. 14, no. 1, pp. 539–545, 2022, doi: 10.1007/s41870-018-0251-8.
[22] Abdullah-All-Tanvir, I. Ali Khandokar, A. K. M. Muzahidul Islam, S. Islam, and S. Shatabda, "A gradient boosting classifier for purchase intention prediction of online shoppers," *Heliyon*, vol. 9, no. 4, 2023, doi: 10.1016/j.heliyon.2023.e15163.
[23] M. Z. Alam and T. Roy, "Predicting Online Repeat Purchases: A Comparative Analysis of Machine Learning Algorithms," *2025 Int. Conf. Electr. Comput. Commun. Eng. ECCE 2025*, 2025, doi: 10.1109/ECCE64574.2025.11013423.
[24] S. xia Chen, X. kang Wang, H. yu Zhang, and J. qiang Wang, "Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine," *Expert Syst. Appl.*, vol. 173, 2021, doi: 10.1016/j.eswa.2021.114756.
[25] Z. Somogyi, "Performance Evaluation of Machine Learning Models," *Appl. Artif. Intell.*, pp. 87–112, 2021, doi: 10.1007/978-3-030-60032-7_3.
[26] G. Varoquaux and O. Colliot, "Evaluating Machine Learning Models and Their Diagnostic Value," *Neuromethods*, vol. 197, pp. 601–630, 2023, doi: 10.1007/978-1-0716-3195-9_20.
[27] A. Tschalzev, S. Marton, S. Lüdtke, C. Bartelt, and H. Stuckenschmidt, "A Data-Centric Perspective on Evaluating Machine Learning Models for Tabular Data," *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, doi: 10.52202/079017-3039.
[28] P. Koukaras and C. Tjortjis, "Data Preprocessing and Feature Engineering for Data Mining: Techniques, Tools, and Best Practices," *AI*, vol. 6, no. 10, 2025, doi: 10.3390/ai6100257.
[29] G. Ravichandra, Bs & Kesavraj, "A Study on Factors Affecting Impulse Buying Behaviour of Apparel Consumer," *J. Xi'an Univ. Archit. Technol.*, vol. 13, no. 1, pp. 515–527, 2021.
[30] N. Singh and A. K. Rai, "Impact Of Influencing Factors On Consumer Buying Behaviour: An Analysis," *Educ. Adm. Theory Pract.*, 2023, doi: 10.53555/kuey.v29i1.7103.
[31] F. Bolikulov, R. Nasimov, A. Rashidov, F. Akhmedov, and Y. I. Cho, "Effective Methods of Categorical Data Encoding for Artificial Intelligence Algorithms," *Mathematics*, vol. 12, no. 16, 2024, doi: 10.3390/math12162553.
[32] Z. Yixuan, "Utilizing machine learning algorithms for consumer behaviour analysis," *Appl. Comput. Eng.*, vol. 49, no. 1, pp. 213–219, 2024, doi: 10.54254/2755-2721/49/20241186.
[33] Y. B. Wah *et al.*, "Machine Learning and Synthetic Minority Oversampling Techniques for Imbalanced Data: Improving Machine Failure Prediction," *Comput. Mater. Contin.*, vol. 75, no. 3, pp. 4821–4841, 2023, doi: 10.32604/cmc.2023.034470.
[34] I. AlShourbaji, N. Helian, Y. Sun, A. G. Hussien, L. Abualigah, and B. Elnaim, "An efficient churn prediction model using gradient boosting machine and metaheuristic optimization," *Sci. Rep.*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-41093-6.
[35] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, [Online]. Available: https://cursa.ihmc.us/rid=1R440PDZR-13G3T80-2W50/4. Pautas-para-evaluar-Estilos-de-Aprendizajes.pdf.
[36] P. Florek and A. Zagdański, "Benchmarking state-of-the-art gradient boosting algorithms for classification," 2023, [Online]. Available: http://arxiv.org/abs/2305.17094.

Vol. 7, No. 2, April 2026, Page. 1441-1454
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5461

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[38] Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., and Gulin A., "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, pp. 6639–6649, 2019.

[39] G. Naidu, T. Zuva, and E. M. Sibanda, "A Review of Evaluation Metrics in Machine Learning Algorithms," *Lect. Notes Networks Syst.*, vol. 724 LNNS, pp. 15–25, 2023, doi: 10.1007/978-3-031-35314-7_2.

[40] A. Odeh, Q. A. Al-Haija, A. Aref, and A. A. Taleb, "Comparative Study of CatBoost, XGBoost, and LightGBM for Enhanced URL Phishing Detection: A Performance Assessment," *J. Internet Serv. Inf. Secur.*, vol. 13, no. 4, pp. 1–11, 2023, doi: 10.58346/JISIS.2023.I4.001.

[41] M. R. Hossain, "Predicting Customer Churn in Telecommunications with Machine Learning Models," *Asian J. Res. Comput. Sci.*, vol. 18, no. 1, pp. 53–66, 2025, doi: 10.9734/ajrcos/2025/v18i1548.

[42] V. Suryanarayana *et al.*, "An efficient implementation of credit card fraud detection using CatBoost algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 38, no. 3, p. 1914, 2025, doi: 10.11591/ijeecs.v38.i3.pp1914-1923.