Applied Information System and Management (AISM)
Volume 8, (1) 2025, p. 103–110
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i1.45025

Enhancing Repeat Buyer Classification with Multi Feature Engineering in Logistic Regression

Siska Farizah Mauludiah[1,2*], Cahyo Crysdian[3], Yunifa Miftachul Arif[4]

***Abstract*—This study presents a novel approach to improving repeat buyer classification on e-commerce platforms by integrating Kullback-Leibler (KL) divergence with logistic regression and focused feature engineering techniques. Repeat buyers are a critical segment for driving long-term revenue and customer retention, yet identifying them accurately poses challenges due to class imbalance and the complexity of consumer behavior. This research uses KL divergence in a new way to help choose important features and evaluate the model, making it easier to understand and more effective at classifying repeat buyers, unlike traditional methods. Using a real-world dataset from Indonesian e-commerce with 1,000 records, divided into 80% for training and 20% for testing, the study uses logistic regression along with techniques like SMOTE for oversampling, class weighting, and regularization to fix issues with data imbalance and overfitting. Model performance is assessed using accuracy, precision, recall, F1-score, and KL divergence. Experimental results indicate that the KL-enhanced logistic regression model significantly outperforms the baseline, especially in balancing precision and recall for the minority class of repeat buyers. The unique contribution of this work lies in its synergistic use of KL divergence in both the feature engineering and evaluation phases, offering a robust, interpreted, and data-efficient solution. For e-commerce businesses, the findings translate into improved targeting of high-value customers, better personalization of marketing efforts, and more strategic allocation of resources. This research offers practical tips for enhancing predictive customer analytics and supports data-driven decision-making in digital commerce environments.**

***Index Terms*—Repeat buyer, classification, logistic regression, feature engineering.**

I. Introduction

In the rapidly evolving e-commerce industry, accurately identifying repeat buyers is essential for improving customer retention, personalizing marketing efforts, and enhancing overall profitability. Repeat buyers, defined as customers who make multiple purchases over a period, are known to exhibit stronger brand loyalty and contribute significantly to a business's revenue through higher lifetime value, making them a key strategic segment for online retailers [1]. The ability to reliably classify such buyers enables e-commerce platforms to allocate marketing budgets more effectively, design personalized promotional campaigns, and implement loyalty programs that resonate with high-value customers. Furthermore, understanding repeat buying behavior supports data-driven decision-making, allowing businesses to anticipate customer needs, reduce churn, and foster long-term relationships. By leveraging advanced analytical techniques for repeat buyer classification, platforms can not only enhance their operational efficiency but also gain a competitive edge in an increasingly saturated digital marketplace [2].

However, predicting repeat purchase behavior is inherently complex due to the diverse and dynamic nature of customer interactions within digital commerce ecosystems. Customers engage with platforms through varied touchpoints and make decisions influenced by a multitude of factors, including demographic characteristics, past transaction patterns, product types, seasonal fluctuations, marketing campaigns, and individual preferences [3]. This variability creates challenges in modeling consistent behavioral patterns, especially when using conventional classification techniques. Traditional models like logistic regression, while interpretable and widely used, often struggle to accommodate such heterogeneity and non-linear relationships when applied in their standard form [4]. To address these limitations, a more sophisticated approach is required, one that leverages the strengths of both linear and non-linear models. By integrating logistic regression with advanced machine learning methods such as XGBoost, it is possible to reduce overfitting while improving model generalization across diverse customer segments. The

Received: 20 February 2025; Revised: 26 April 2025; Accepted: 10 May 2025
*Corresponding author
[1]*Siska Farizah Mauludiah, Master Study of Computer Science, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia (e-mail: siskafm@yahoo.com).
[2]Siska Farizah Mauludiah, PT. Theta Memontum Information Technology, Indonesia (e-mail: siskafm@yahoo.com).
[3]Cahyo Crysdian, Master Study of Computer Science, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia (e-mail: cahyo@ti.uin-malang.ac.id)
[4]Yunifa Miftachul Arif, Department of Informatics Engineering, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia (e-mail: yunif4@ti.uin-malang.ac.id).

ensemble capabilities of XGBoost, when combined with techniques like K-fold cross-validation, allow for better handling of data variance and improve the accuracy of repeat purchase predictions across multiple validation sets [5].

Logistic regression remains widely used for binary classification tasks due to its simplicity, interpretability, and computational efficiency, especially when applied to large-scale datasets commonly encountered in e-commerce applications [6]. Its clear mathematical foundation and straightforward implementation make it a preferred baseline model for many predictive analytics tasks. However, despite these advantages, logistic regression often encounters performance bottlenecks when faced with real-world data complexities. For instance, the model's assumptions about linear relationships between features and the target variable can limit its effectiveness in capturing underlying patterns within customer behavior data. Additionally, it is particularly sensitive to imbalanced class distributions, where the minority class (e.g., repeat buyers) is underrepresented, leading to biased predictions and poor recall [7]. Furthermore, high-dimensional feature spaces, resulting from detailed customer attributes, transaction records, and encoded categorical variables, can introduce noise and multicollinearity, further reducing model stability and predictive power. Addressing these challenges necessitates a more nuanced and adaptive modeling framework that incorporates advanced feature engineering strategies, such as dimensionality reduction, variable selection, and distribution-aware transformations, alongside robust evaluation methods to ensure fairness and accuracy [8].

Classification algorithms integrated with feature selection techniques can effectively address critical challenges such as class imbalance, multicollinearity among features, and high-dimensional data structures that are typical in real-world applications [9]. These techniques help refine the feature space by removing redundant or irrelevant variables, thereby improving model interpretability and reducing the risk of overfitting. In this study, the methodology is applied to a real-world e-commerce dataset from Indonesia, which captures a wide array of variables including transaction date, customer ID, age, gender, geographic area, product ID, product name, product category, and purchase amount. This comprehensive dataset provides a valuable basis for understanding customer behavior in a localized market context. By analyzing both customer purchasing history and demographic information, the model can identify patterns indicative of repeat buying tendencies. These indicators serve as proxies for deeper behavioral attributes such as customer loyalty, trust, and satisfaction. Consequently, the resulting insights not only enhance the accuracy of repeat buyer classification but also support more targeted and effective customer monitoring, segmentation, and relationship management strategies within the company's operational framework [10].

Central to the proposed framework is the application of Kullback-Leibler (KL) divergence, which quantifies the difference between probability distributions and helps prioritize features based on their information gain [11]. Moreover, the integration of SMOTE and feature engineering in refining repeat buyer prediction can enhance the model's AUC score [12], while advanced machine learning techniques are capable of achieving high validation accuracy by effectively capturing complex data patterns [13].

By applying KL divergence in feature selection, the model is better equipped to differentiate between repeat and one-time buyers. In addition to KL divergence, the feature engineering pipeline incorporates mutual information, correlation analysis, histogram-based binning, and dimensionality reduction techniques to improve data quality and reduce redundancy [14]. Logistic regression is further enhanced with class weighting and regularization techniques to prevent overfitting and support better generalization across diverse buyer segments [15].

The classification model is evaluated using standard performance metrics including accuracy, precision, recall, and F1-score, which provide a balanced assessment of predictive reliability [16]. By integrating these metrics, the model ensures consistent performance even under skewed class distributions and sparse data environments [17]. Furthermore, the optimal predictive solution, as indicated by metrics such as accuracy, precision, recall, and F1-score, enables the identification of the most suitable predictive model while addressing the limitations of traditional evaluation methods by emphasizing balanced accuracy and context-aware performance measurement [18].

The combination of KL divergence-based feature engineering, class balancing techniques, and logistic regression forms a powerful and interpretable model for repeat buyer classification [19]. This model supports actionable insights for customer segmentation and strategic marketing decisions, offering scalable benefits to e-commerce platforms seeking to increase retention and revenue [20]. The integration of synthetic oversampling, feature selection, and performance refinement reflects a practical, data-driven approach to overcoming existing limitations in customer behavior prediction [21].

Ultimately, an effective resolution to the challenge of repeat buyer identification can be achieved by enhancing classification performance through strategic feature engineering and the application of adaptive modeling techniques [22]. Techniques such as SMOTE and ADASYN enhance classifier performance [23], while CONMI_FS filters for optimal feature subset selection can yield robust results with KNN and SVM [24]. Building on these approaches, a systematic simulation-based analysis of performance metrics for imbalanced datasets has identified the Matthews Correlation Coefficient as the most suitable metric when classification errors are critical, while also proposing new class balance metrics that extend beyond traditional accuracy measures [25]. The implementation of robust performance metrics such as F1-score and MCC adds further reliability to the evaluation process, ensuring that both prediction quality and error distribution are meaningfully captured, especially in scenarios where class imbalance is a significant concern [26].

By building upon logistic regression and enhancing it with advanced feature engineering techniques, the incorporation of KL divergence enables the identification of the most relevant features within a dataset. The KL divergence approach has been shown to improve classification accuracy by effectively guiding feature selection during the feature extraction process [27]. An analysis of logistic regression and other machine learning algorithms in predicting repeat buyers highlights their

Applied Information System and Management (AISM)
Volume 8, (1) 2025, p. 103–110
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i1.45025

performance under imbalanced class distributions. By critically evaluating the limitations of the models, the approach integrates feature engineering, Synthetic Minority Over-sampling Technique (SMOTE), and hyperparameter tuning to enhance predictive accuracy and effectively address class imbalance [28]. Moreover, the inclusion of diverse consumption indicators and ensemble modeling enhances the model's predictive power, while comparative metric evaluations validate the strength of logistic regression in real-world applications [29].

The objectives of this study are as follows:

- To enhance repeat buyer classification by integrating KL divergence into the feature engineering process. This involves using KL divergence to identify and retain features that exhibit the most significant distributional differences between repeat and non-repeat buyers. By focusing on these high-information features, the model is expected to improve its ability to distinguish between the two classes effectively.
- To apply logistic regression augmented with SMOTE, class weighting, and regularization for improved model performance. SMOTE is used to address class imbalance by oversampling the minority class (repeat buyers), class weighting adjusts the model's sensitivity to misclassification, and regularization helps prevent overfitting. Together, these techniques aim to create a more balanced, robust, and generalizable model.
- To evaluate the model using a real-world e-commerce dataset through standard performance metrics. The model's effectiveness will be assessed using accuracy, precision, recall, and F1-score, alongside KL divergence, to ensure comprehensive performance evaluation on practical data collected from actual customer transactions.
- To offer interpretable and practical insights for improving customer segmentation and retention strategies in e-commerce. By developing a model that is both transparent and reliable, the study aims to provide actionable outcomes that can assist e-commerce businesses in identifying valuable customers and tailoring strategies to enhance long-term engagement and profitability.

II. Related Work

A. Repeat Buyer Classification

Repeat buyer classification is a vital focus in e-commerce, enabling businesses to identify customers who are likely to make multiple purchases and thereby contribute to long-term profitability. These buyers tend to exhibit loyalty, higher engagement, and more predictable behavior patterns [2]. In contrast, one-time purchasers often respond primarily to short-term discounts or incentives, making them less valuable in the long run. By identifying repeat buyers early, companies can implement more targeted marketing and retention strategies [1], [6]. Past studies have emphasized the importance of understanding user activity patterns and consumer decision-making behaviors, noting that e-commerce data, such as purchase frequency, recency, and product preferences, offers critical insights into customer segmentation [5]. However, several challenges still hinder effective classification, including data sparsity, class imbalance, and inconsistent feature relevance [12]. In recent years, the availability of large-scale behavioral data from online platforms has fueled more nuanced approaches to classification. This includes leveraging browsing activity and purchase history to model future buying intent [17], while also applying structured feature engineering techniques to extract relevant signals from raw data [16].

This study builds upon these efforts by emphasizing the need to refine and select informative features from transactional and demographic data. It seeks to improve the accuracy and robustness of repeat buyer prediction models, especially in the face of data sparsity and high variability in consumer behavior.

B. Classification Models

Numerous machine learning models have been utilized for binary classification tasks in e-commerce, including Support Vector Machines (SVM), Decision Trees, Naïve Bayes, Artificial Neural Networks, K-Nearest Neighbors, and Random Forests [7], [10], [24]. These models offer varying trade-offs between interpretability, performance, and scalability. While ensemble models like Random Forest and neural networks can capture complex interactions and improve predictive accuracy, they often lack transparency and may require extensive tuning.

Logistic regression, in contrast, is widely favored for its simplicity and ease of interpretation [3], [6]. It has been applied effectively in predicting repeat purchases, particularly when models must remain understandable to marketing and business stakeholders [27]. However, logistic regression's linear assumptions limit its ability to capture non-linear relationships among features, especially when faced with imbalanced datasets or high-dimensional feature spaces [4]. This study adopts logistic regression as the base model but enhances it with advanced feature engineering and class balancing techniques to overcome these limitations while maintaining interpretability.

C. Kullback-Leibler Divergence

Kullback-Leibler (KL) divergence has emerged as a valuable method for feature selection, especially when dealing with high-dimensional data. As a divergence measure, it quantifies how different the distribution of a feature is between classes, making it suitable for identifying features that most effectively distinguish between repeat and one-time buyers [8], [11]. Unlike wrapper or embedded methods, filter-based techniques like KL divergence are computationally efficient and work independently of the classification algorithm. This efficiency is especially beneficial in early-stage model development or when computational resources are limited.

KL divergence has been shown to reduce dimensionality effectively, improving texture classification and other machine learning tasks by eliminating noisy and irrelevant features [14], [26]. By selecting only the most representative variables, models can train faster and predict more accurately. In this study, KL divergence is used both as a feature selection method

and as an evaluation metric to assess feature quality, providing a consistent framework for refining input variables throughout the classification pipeline.

D. Feature Engineering and Imbalanced Data

Feature engineering is a cornerstone of effective machine learning, particularly in domains like e-commerce, where raw data can be noisy, sparse, and high-dimensional. Creating or selecting meaningful features helps models better capture underlying patterns and improve predictive accuracy. Despite preprocessing efforts, challenges such as class imbalance and strong inter-feature correlations often persist [18], [20]. These issues can distort model training and lead to misleading performance evaluations if not properly addressed [19].

To overcome these problems, recent studies have combined classification models with balancing methods such as the Synthetic Minority Over-sampling Technique (SMOTE) [9], [22]. Feature selection techniques that combine Pearson correlation coefficients and normalized mutual information have also been introduced to capture both linear and non-linear dependencies between features and target variables [23]. This ensemble-style feature selection enables more robust modeling of complex consumer behavior.

In alignment with these approaches, the present study applies multiple feature engineering techniques—including KL divergence thresholding, mutual information analysis, correlation filtering, and bin-based histogram computation—to select and refine features. Furthermore, SMOTE, class-weighted logistic regression, and regularization are integrated into the modeling process to address data imbalance and enhance the generalizability of the logistic regression classifier.

III. Research Method

This study employed a structured methodology to enhance repeat buyer classification using logistic regression combined with Kullback-Leibler (KL) divergence and multi-feature engineering techniques. The research followed four key stages: Data Preparation, Feature Engineering, Model Training, and Evaluation. Each stage is explained in the next subsections. Figure 1 illustrates the complete workflow of the proposed method.

A. Data Preparation

An e-commerce transaction dataset was obtained from a digital transformation service company in Indonesia. The dataset comprises the fields listed in Table 1.

The preparation phase involved cleaning and preprocessing customer purchase history and demographics. The transaction date (Trx_Date) was converted to datetime format, and a new feature (YearMonth) was extracted. Transactions were grouped by Cust_ID to derive aggregated features, including the number of transactions, first and last transaction dates, number of active months, total and average amount spent, and number of distinct product categories. A binary label, Repeat_Buyer, was introduced to indicate whether the customer transacted in more than one month. Demographic variables (Age, Gender, Area) were merged to form a customer-level dataset as shown in Table 2.

Fig 1. Workflow of proposed method.

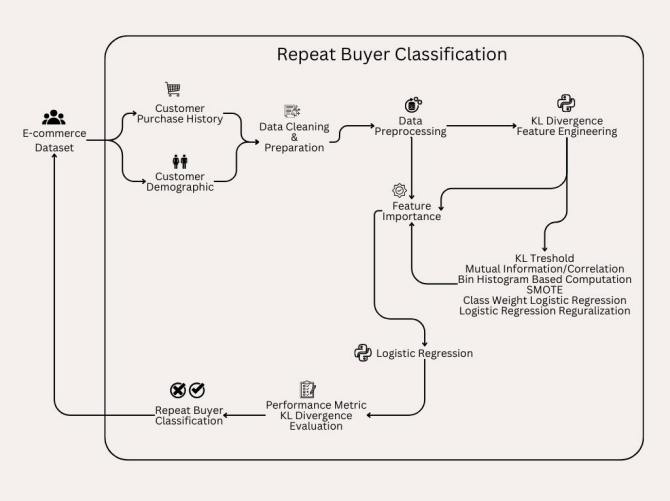

Table 1.
Fields in The Dataset

Field	Description
Trx_Date	Transaction Date
Cust_ID	Customer ID
Age	Customer Age
Gender	Customer Gender
Area	Customer Location
Product_ID	Product ID
Product_Name	Product Name
Category	Product Category
Amount	Transaction Amount

Table 2.
Final Data Attributes

Attribute	Description
Cust_ID	Unique identifier for each customer
First_trx	Date of the customer's first recorded transaction
Last_trx	Date of the customer's most recent recorded transaction
Trx_count	Total number of transactions made by the customer
Distinct_Periods	Number of distinct months the customer made a purchase
Total_Quantity	Total amount of items purchased by the customer
Avg_Quantity	Average number of items purchased per transaction
Unique_Categories	Number of different product categories purchased
Repeat_Buyer	Binary label indicating whether the customer is a repeat buyer (1) or not (0)
Age	Age of the customer
Gender	Gender of the customer
Area	Geographic area or region where the customer resides

B. Feature Engineering

Three experimental setups were defined.

1) Baseline Logistic Regression

Applied Information System and Management (AISM)
Volume 8, (1) 2025, p. 103–110
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i1.45025

The baseline model directly applied logistic regression to the full dataset. The logistic function calculates the probability that a customer is a repeat buyer as the formula in (1).

$$P(y=1 \mid X) = \frac{1}{1+e^{(-\beta_0+\beta_1 X_1+\ldots+\beta_n X_n)}} \quad (1)$$

where β_0 is the intercept and β_1 through β_n are coefficients for input features X_1 to X_n. Predictions above 0.5 were classified as repeat buyers (Class 1).

2). KL Divergence-Based Feature Selection

KL divergence was used to measure how differently a feature is distributed between repeat and non-repeat buyers. Features with high KL divergence values were selected for model training, indicating greater discriminative power. This technique helped reduce noise and focus on informative attributes. As shown in (2), it calculates the probability distribution for each feature x and then calculates the KL divergence.

$$D_{KL}(P \| Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \quad (2)$$

3). Multi-Feature Engineering

KL divergence was integrated with other techniques to improve performance:

a. KL Threshold Adjustment: The current threshold for feature selection could be too high or too low, leading to over-simplification of irrelevant features. Thresholds (0.05, 0.1, 0.2) were tested to filter weak features.
b. Mutual Information & Correlation Analysis: Features were ranked by mutual information gain and reduced for multicollinearity. By selecting features which provide the greatest information gain, mutual information aids with feature selection by quantifying the dependence between variables. Recurrent features that could cause multicollinearity in the model can be found using correlation analysis. Steps to combining KL divergence with mutual information (MI) or correlation to ensure the feature set captured both class separability and relevance are computed mutual information for comparison, and then combined KL divergence with MI scores. The scores was normalized for weighted combination and combine the cores to select the top features.
c. Histogram Binning: Features were binned (10-20 bins) to better capture consumer behavior. Employing this method, continuous variables are divided into bins (such as purchase frequency ranges), and distributional patterns are captured by computing histograms. These histogram-based features reveal patterns in consumer behavior.
d. SMOTE: Used to oversample the minority class (repeat buyers). To enhance model performance and avoid bias toward the majority class, SMOTE generates synthetic examples of the minority class when there is a class imbalance, meaning that one class has substantially less data than the other.
e. Class-Weighted Logistic Regression: Applied penalty for misclassifying minority class. Assigning distinct weights to various classes in logistic regression, as opposed to resampling, aids in managing unbalanced datasets by harshly penalizing incorrect classifications of underrepresented groups.
f. Regularization (L1, L2): Prevented overfitting by constraining model complexity. To minimize overfitting in logistic regression models, L1 (Lasso) and L2 (Ridge) regularization techniques are applied. While L2 regularization aids in controlling significant coefficients to enhance model generalization, L1 regularization may result in sparse models by eliminating less significant features. The hyperparameter grid was defined by smaller c to 0.01, 01 and larger c to 10,100. c parameter controls the inverse of regularization strength in logistic regression.

4). Model Training

To evaluate the effectiveness of different approaches in classifying repeat buyers, experimental configurations of logistic regression were conducted as follows:

First, the baseline model was trained using the raw, unprocessed features directly from the dataset. This initial setup serves as a control model to understand how well logistic regression performs without any enhancements, providing a reference point for further comparisons.

Second, the model was improved by applying feature selection using KL-divergence, where only the most informative features, those showing the greatest distributional differences between repeat and non-repeat buyers, were retained. This step aimed to reduce noise and focus the model on variables most relevant to the classification task, potentially improving accuracy and interpretability.

Third, the most advanced configuration combined KL-divergence-selected features with additional feature engineering techniques, including Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance, and regularization to prevent overfitting. This comprehensive approach aimed to maximize model generalization while ensuring fair treatment of the minority class (repeat buyers).

All three models were trained and tested using the same dataset split to ensure a fair comparison. This experimental design enabled the assessment of incremental performance gains at each level of feature refinement and model enhancement, guiding the selection of the most effective configuration for repeat buyer classification.

5). Evaluation

The models were evaluated using a comprehensive set of performance metrics, including accuracy, precision, recall, and F1-score, to assess their ability to correctly classify repeat buyers. These metrics provided a balanced view of

performance, particularly in addressing class imbalance. In addition to standard evaluation, Kullback-Leibler (KL) divergence was employed to compare the distributions of actual and predicted probabilities, offering a deeper insight into how well the model's output aligned with the true data distribution.

To further validate the effectiveness of feature engineering, KL divergence scores were calculated before and after applying advanced techniques. The results showed a significant reduction in divergence, indicating that the engineered features enhanced the model's ability to distinguish between repeat and non-repeat buyers. This confirmed that the improvements were not only reflected in traditional metrics but also in the underlying probabilistic structure of the model, supporting the overall robustness of the enhanced logistic regression approach.

IV. RESULT

The results of the classification experiments are presented in four stages, beginning with the baseline logistic regression, followed by KL divergence-based feature selection, multi-feature engineering integration, and concluding with comparative performance and KL divergence-based evaluation. Each stage highlights the main outcomes and transitions into the next step to explain how the model evolved and improved.

The baseline logistic regression model initially identified AREA and Distinct_Periods as the most influential predictors of repeat buyer behavior, as shown in Fig 2.

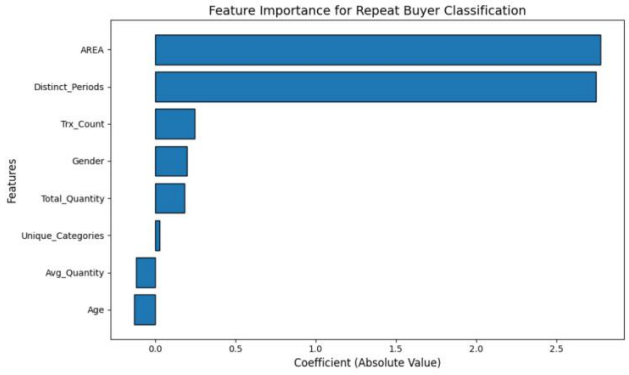

Fig. 2. Logistic regression feature importance.

Although the model demonstrated strong performance with an accuracy of 90% as can be seen in Table 3, further analysis revealed signs of overfitting and data triviality. The performance metrics appeared nearly perfect across precision, recall, and F1-score, suggesting the model may have over-learned from easily distinguishable patterns in the dataset. To mitigate this issue, the model was recalibrated using regularization (L1 and L2), SMOTE for class balancing, and mutual information-based feature selection.

Table 3.
Logistic Regression Performance Metrics

Class	Precision	Recall	F1-Score
0	0.94	0.89	0.91
1	0.85	0.92	0.88
Accuracy		0.90	
Macro Avg	0.90	0.90	0.90
Weighted Avg	0.90	0.90	0.90

Following this, KL divergence was applied to evaluate feature distributions between repeat and non-repeat buyers. This method highlighted Total_Quantity and Unique_Categories as more discriminative features than those identified in the baseline model, as illustrated in Fig 3. Although this improved feature interpretability, model performance showed continued imbalance, particularly in minority class detection.

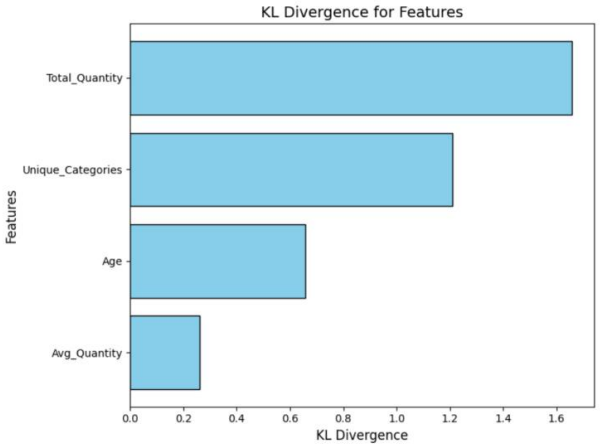

Fig. 3. Feature importance by KL divergence.

And also as indicated in Table 4, recall for repeat buyers (Class 1) was only 0.10, despite perfect precision. This imbalance emphasized the need for a more comprehensive feature engineering strategy.

Table 4.
KL Divergence-Based Features Result

Class	Precision	Recall	F1-Score	Support
0	0.75	1.00	0.86	27
1	1.00	0.10	0.18	10
Accuracy			0.76	37
Macro Avg	0.88	0.55	0.52	37
Weighted Avg	0.82	0.76	0.67	37

To address this limitation, a multi-feature engineering approach was implemented. This strategy combined KL divergence threshold tuning, mutual information and correlation filtering, bin-based histogram computation, SMOTE, class weighting, and regularization. The integrated approach produced marked improvements: recall for repeat buyers increased from 0.10 to 0.30, and F1-score rose from 0.18 to 0.46. Overall accuracy also improved from 76% to 81%, as shown in Table 5. These gains demonstrate that integrating multiple feature engineering techniques enhances the model's capacity to recognize minority class patterns while maintaining performance for the majority class.

Table 5.
KL Divergence with Multi Feature Engineering

Applied Information System and Management (AISM)
Volume 8, (1) 2025, p. 103–110
P-ISSN: 2621-2536; E-ISSN: 2621-2544; DOI: 10.15408/aism.v8i1.45025

Class	Precision	Recall	F1-Score	Support
0 (Majority)	0.79	1.00	0.89	27
1 (Minority)	1.00	0.30	0.46	10
Accuracy			0.81	37
Macro Avg	0.90	0.65	0.67	37
Weighted Avg	0.85	0.81	0.77	37

A comparison of key performance metrics between the baseline logistic regression model and the KL divergence-enhanced model is presented in Table 6. The KL-enhanced model achieved a substantially higher mean F1-score (0.6416 vs. 0.3064) and demonstrated better balance in precision and recall, particularly for the minority class. Although there were slight trade-offs, such as a decrease in precision for Class 1, these were acceptable given the substantial gains in recall and overall classification effectiveness.

Table 6.
Key Metrics Comparison

Metrics	Baseline Logistic Regression	KL Divergence Logistic Regression
Mean F1-Score	0.3064	0.6416
Accuracy	76%	81%
Precision (Class 0)	75%	81%
Precision (Class 1)	100%	80%
Recall (Class 0)	100%	96%
Recall (Class 1)	10%	40%
F1-Score (Class 0)	0.86	0.88
F1-Score (Class 1)	0.18	0.53

Finally, KL divergence was employed not only as a feature selection method but also as an evaluation metric. By measuring how well the predicted class distributions aligned with actual distributions, this metric provided insight into the separability of the selected features. As presented in Table 7, Total_Quantity exhibited a significant increase in KL divergence from 1.6582 to 11.3966 after applying multi-feature engineering. Avg_Quantity also improved from 0.2619 to 3.1449, reinforcing its relevance. In contrast, features like Unique_Categories and Age showed lower post-engineering scores and were excluded from the final model.

Table 7.
KL Divergence Evaluation Result

Feature	Before Applyig Multi Feature Engineering	After Applying Multi Feature Engineering
Total_Quantity	1.6582	11.396643
Unique_Categories	1.2108	-
Age	0.6585	-
Avg_Quantity	0.2619	3.144888

These findings confirm that integrating KL divergence into both the feature selection and evaluation stages significantly enhances the reliability and interpretability of logistic regression models for repeat buyer classification. By identifying and retaining the most informative features, the model becomes more focused and effective, reducing noise and improving predictive accuracy.

Moreover, the comprehensive feature engineering approach, including KL-based selection, SMOTE for class balancing, and regularization, proved especially effective in addressing class imbalance, a common challenge in real-world e-commerce data. This not only led to improved classification performance but also ensured the model's outputs are more aligned with business goals, making it a practical solution for customer retention strategies.

V. Conclusion

This study establishes a robust and interpretable pipeline for repeat buyer classification by integrating logistic regression with Kullback-Leibler (KL) divergence and multi-feature engineering. The results confirm that combining KL divergence with feature selection techniques such as mutual information, correlation analysis, and bin-based computation significantly improves classification performance, especially for the minority class.

The initial baseline logistic regression model struggled with class imbalance, showing signs of overfitting and limited generalizability. By contrast, the enhanced model, leveraging engineered features and balancing strategies, achieved improved recall and F1-score for repeat buyers, thereby increasing the model's ability to detect high-value customers more reliably. These findings underscore the value of KL divergence not only as a feature selection method but also as a diagnostic tool for evaluating class separability.

Beyond improving metrics, this study highlights important practical implications. Businesses operating in the e-commerce domain can benefit from adopting this framework to more accurately identify repeat buyers, allowing for better-targeted retention strategies and resource allocation. The methodology balances performance and interpretability, which is essential for real-world deployment where transparency is often a key requirement.

Future work can further refine this pipeline by incorporating ensemble methods such as XGBoost or LightGBM to compare model performance across architectures. Research may also explore real-time and adaptive resampling strategies that evolve with customer behavior, or test deep learning models with attention mechanisms for more complex datasets. Additionally, integrating divergence-based feature selection into automated machine learning (AutoML) platforms could streamline hyperparameter tuning and improve deployment efficiency.

In conclusion, the integration of KL divergence and multi-feature engineering into a logistic regression framework offers a practical and scalable approach for imbalanced classification problems, especially those in customer analytics and digital commerce.

Acknowledgment

Thanks to all members of Master Study of Computer Science, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia for great support during the study and the

research.

REFERENCES

[1] K. Kareena and R. Kumar, "A consumer behavior prediction method for e-commerce application," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2 Special Issue 6, pp. 983–988, Jul. 2019, doi: 10.35940/ijrte.B1171.0782S619.

[2] T. Charanasomboon and W. Viyanon, "A comparative study of repeat buyer prediction: Kaggle acquired value shopper case study," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, 2019, pp. 306–310. doi: 10.1145/3322645.3322681.

[3] C. J. Liu, T. S. Huang, P. T. Ho, J. C. Huang, and C. T. Hsieh, "Machine learning-based e-commerce platform repurchase customer prediction model," *PLoS ONE*, vol. 15, no. 12, pp.1–15, Dec. 2020, doi: 10.1371/journal.pone.0243105.

[4] A. Ahmed, A. Jalal, and K. Kim, "A novel statistical method for scene classification based on multi-object categorization and logistic regression," *Sensors (Switzerland)*, vol. 20, no. 14, pp. 1–20, Jul. 2020, doi: 10.3390/s20143871.

[5] P. Song and Y. Liu, "An xgboost algorithm for predicting purchasing behaviour on e-commerce platforms," *Tehnicki Vjesnik*, vol. 27, no. 5, pp. 1467–1471, Oct. 2020, doi: 10.17559/TV-20200808113807.

[6] H. Zhang and J. Dong, "Prediction of repeat customers on e-commerce platform based on blockchain," *Wireless Communications and Mobile Computing*, vol. 2020, pp.1–15, Aug. 2020, doi: 10.1155/2020/8841437.

[7] B. Noori, "Classification of customer reviews using machine learning algorithms," *Applied Artificial Intelligence*, vol. 35, no. 8, pp. 567–588, 2021, doi: 10.1080/08839514.2021.1922843.

[8] S. Pourmand, A. Shabbak, and M. Ganjali, "Feature selection based on divergence functions: A comparative classification study," *Statistics, Optimization and Information Computing*, vol. 9, no. 3, pp. 587–606, 2021, doi: 10.19139/soic-2310-5070-1092.

[9] T. Wang, P. Chen, T. Bao, J. Li, and X. Yu, "Arrhythmia classification algorithm based on SMOTE and feature selection," *International Journal of Performability Engineering*, vol. 17, no. 3, pp. 263–275, Mar. 2021, doi: 10.23940/ijpe.21.03.p2.263275.

[10] Y. Suhanda, L. Nurlaela, I. Kurniati, A. Dharmalau, and I. Rosita, "Predictive analysis of customer retention using the random forest algorithm," *TIERS Information Technology Journal*, vol. 3, no. 1, pp. 35–47, Jun. 2022, doi: 10.38043/tiers.v3i1.3616.

[11] T. K. Nguyen, Z. Ahmad, and J. M. Kim, "A deep-learning-based health indicator constructor using kullback–leibler divergence for predicting the remaining useful life of concrete structures," *Sensors*, vol. 22, no. 10, Art. no. 3687, May 2022, doi: 10.3390/s22103687.

[12] 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," *International Journal of Computers, Communications and Control*, vol. 17, no. 6, pp.1–17, Dec. 2022, doi: 10.15837/ijccc.2022.6.4988.

[13] M. B. Tamam, H. Hozairi, M.Walid, and J. F. A. Bernardo, "Classification of sign language in real time using convolutional neural network," *Applied Information System and Management (AISM)*, vol. 6, no. 1, pp. 39–46, Apr. 2023, doi: 10.15408/aism.v6i1.29820.

[14] F. Mendonça, S. S. Mostafa, F. Morgado-Dias, and A. G. Ravelo-García, "On the use of kullback–leibler divergence for kernel selection and interpretation in variational autoencoders for feature creation," *Information (Switzerland)*, vol. 14, no. 10, pp.1–15, Oct. 2023, doi: 10.3390/info14100571.

[15] U. Buatoom and M. U. Jamil, "Improving classification performance with statistically weighted dimensions and dimensionality reduction," *Applied Sciences (Switzerland)*, vol. 13, no. 3, pp.1–20, Feb. 2023, doi: 10.3390/app13032005.

[16] R. Kc, S. Shandilya, and M. Shandilya, "Unlocking Future Transactions: Predicting Customer's Next Purchase in E-commerce through Machine Learning Analysis," *International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIE)*, vol. 9, no. 3, pp. 1077–1081, Apr. 2023.

[17] L. Li, "Analysis of e-commerce customers' shopping behavior based on data mining and machine learning," *Soft Computing*, vol. 27, no. 29, pp. 1–14, Jul. 2023, doi: 10.1007/s00500-023-08903-5.

[18] M. Owusu-Adjei, J. Ben Hayfron-Acquah, T. Frimpong, and G. Abdul-Salaam, "Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems," *PLOS Digital Health*, vol. 2, no. 11, pp. 1–19, Nov. 2023, doi: 10.1371/journal.pdig.0000290.

[19] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced Big Data," *Journal of Big Data*, vol. 10, no. 1, pp.1–31, Dec. 2023, doi: 10.1186/s40537-023-00724-5.

[20] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah. "Comparative analysis using various performance metrics in imbalanced data for multi-class text classification," *International Journal of Data Science,* vol. 5 no.2, pp. 45–53. July 2023. doi:10.14569/IJACSA.2023.01406116.

[21] D. Dablain, B. Krawczyk, and N. v. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: 10.1109/TNNLS.2021.3136503.

[22] M. Mujahid, E. Kına, F. Rustam, M. G. Villar, E. S. Alvarado, I. D. L. T. Díez, and I. Ashraf, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *Journal of Big Data*, vol. 11, no. 1, pp.1–32, Dec. 2024, doi: 10.1186/s40537-024-00943-4.

[23] H. Gong, Y. Li, J. Zhang, B. Zhang, and X. Wang, "A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information," *Engineering Applications of Artificial Intelligence*, vol. 131, Art. no. 107865, May. 2024, doi: 10.1016/j.engappai.2024.107865.

[24] D. C. Gkikas and P. K. Theodoridis, "Predicting online shopping behavior: Using machine learning and google analytics to classify user engagement," *Applied Sciences (Switzerland)*, vol. 14, no. 23, pp.1-31, Dec. 2024, doi: 10.3390/app142311403.

[25] A. Luque, A. Carrasco, A. Martín, and A. D. L. Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, May. 2019, doi: 10.1016/j.patcog.2019.02.023.

[26] M. S. Rao *et al.*, "Kullback–leibler divergence-based feature selection method for image texture classification," in *Lecture Notes in Networks and Systems*, Springer Science and Business Media Deutschland GmbH, 2024, pp. 309–318. doi: 10.1007/978-981-99-9704-6_27.

[27] S. F. Mauludiah, Y. M. Arif, M. Faisal, and D. D. Putra, "Struggling models: an analysis of logistic regression and random forest in predicting repeat buyers with imbalanced performance metrics," *Applied Information System and Management (AISM)*, vol. 7, no. 2, pp. 31–38, Jul. 2024, doi: 10.15408/aism.v7i2.39326.

[28] N. Yue, "Identify potential loyalists in shopping festival: repeat buyer prediction for e-commerce based on feature engineering and ensemble learning," M.S. thesis, Erasmus School of Economics, Erasmus Univ. Rotterdam, Rotterdam, Netherlands, Sept. 2024. [Online]. Available: https://thesis.eur.nl/pub/72505

[29] P. A. Sunarya, U. Rahardja, S. C. Chen, Y. M. Li, and M. Hardini, "Deciphering digital social dynamics: A comparative study of logistic regression and random forest in predicting e-commerce customer behavior," *Journal of Applied Data Sciences*, vol. 5, no. 1, pp. 100–113, Jan. 2024, doi: 10.47738/jads.v5i1.155.