E-ISSN : 3025-1311
https://journal.cib.institute/index.php/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 11 No 5 Tahun 2026

# EVALUATING BEHAVIORAL AND TEMPORAL FEATURES FOR E-COMMERCE PURCHASE PREDICTION: A COMPARISON OF LOGISTIC REGRESSION AND CATBOOST

**Aghasi Yanda Wafa Azizah[1*], Naza Sulthoniyah Wahda[2], Putri Ikrima Al Fadhilah[3], Ryan Londhut Nathaniel[4], Yahya Dai Bahtiar[5]**

[1-3]Data Science, Universitas Negeri Surabaya
[4]Renewable Energies, Hochschule für Technik und Wirtschaft Berlin
[5]Informatics Engineering, Universitas PGRI Madiun

*Email: aghasi.23039@mhs.unesa.ac.id

**ABSTRACT**

Purchase prediction on e-commerce platforms can help identify the likelihood of consumers completing a purchase after adding a product to the shopping cart. This study aims to evaluate the use of behavioral and temporal features to predict the purchase of the same product within 24 hours of add-to-cart activity by comparing the Logistic Regression and CatBoost algorithms. The study uses secondary data from the Synerise RecSys Challenge 2025 which includes product addition to cart activity and purchase activity. A total of 149,208 observations were used in modeling by dividing data based on time order into training, validation, and test data. The features used included the frequency of cart activity, user purchase history, time interval since previous activity or purchase, and temporal characteristics based on calendar. Model evaluation was conducted using ROC-AUC, Average Precision (AP), accuracy, precision, recall, and F1-score. The test results showed that CatBoost outperformed Logistic Regression on all six evaluation metrics, with a ROC-AUC of 0.6830, an AP of 0.4212, an accuracy of 0.6926, a precision of 0.3820, a recall of 0.5530, and an F1-score of 0.4518. A feature significance analysis showed that the number of users' historical purchases, the frequency of cart activity, and the time interval since the user's last purchase were the features that contributed the most to the prediction. In addition, the addition of customer identities and SKUs did not increase validation AP compared to the use of numerical features alone. The results showed that historical behavioral information has the potential to support short-term purchase predictions and that CatBoost provided better performance in the experiments conducted. However, testing on additional periods and datasets is still needed to assess the generalization capabilities and application of the model under real conditions.

**Keywords:** E-commerce; purchase prediction; behavioral features; temporal features; logistic regression; CatBoost.

**Article History**
Received: Oktober 2026
Reviewed:Oktober 2026
Published: Oktober 2026

Plagiarism Checker No 237.u873247.fh8374

**Copyright : Author**
**Publish by : Kohesi**

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

## Introduction

The development of e-commerce has led to an increase in the amount of consumer interaction data that can be used to understand purchasing behavior and support data-driven decision-making. User activities, such as viewing products, adding products to the cart (add-to-cart), and completing transactions, reflect the stages in the online purchase process. However, the addition of products to the cart is not always followed by transactions. Chopra et al. (2024) explain that the phenomenon of shopping cart abandonment is influenced by various factors related to consumers and the online shopping experience. This condition shows the importance of developing a predictive approach to identify cart activities that have the potential to result in purchases so that they can support marketing strategies and manage customer interactions in a more targeted manner.

The use of *machine learning* is one of the approaches that is developing in the analysis of consumer behavior on e-commerce platforms. Wen et al. (2023) developed a purchase intent prediction approach using clickstream data by considering user behavior information and product characteristics. Meanwhile, Alamuri and Bondalapu (2026) examined the application of machine learning algorithms to predict consumer purchase intent. Both studies show that digital activity patterns can provide relevant information for the development of prediction models. However, the prediction of purchase intent based on session activity is not entirely the same as the prediction of actual transactions for a particular product within a predetermined time horizon.

In addition to the characteristics of the data, the selection of algorithms affects the model's ability to recognize purchasing patterns. Logistic Regression is one of the classification algorithms that can be used as a baseline model because it has a relatively simple structure and allows the interpretation of the relationship between predictor variables and positive class probabilities. In contrast, CatBoost is a gradient boosting-based algorithm that is able to study nonlinear relationships and interactions between features. Kafilla et al. (2025) evaluated the ensemble machine learning approach in predicting consumer purchasing behavior, including comparisons with Logistic Regression. The study shows the relevance of the evaluation of various classification approaches, but differences in dataset characteristics, target definitions, and test procedures mean that the results between studies cannot be directly compared.

Feature representation is also an important component in modeling purchasing behavior. Information about interaction frequency, transaction history, and activity recency can describe the intensity and historical behavior patterns of users. In addition, temporal characteristics, such as hours, days, and months of activity, have the potential to represent variations in interaction based on time. Wen et al. (2023) show the use of user behavior information in predicting purchase intent based on clickstream data. Bastos and Bernardes (2024) also examined online purchase predictions using explainable machine learning approaches to support the interpretation of factors related to prediction outcomes. The study provides a foundation for the use of behavioral features and feature contribution analysis, although the benefits of each feature group still need to be tested based on the characteristics of the prediction task used.

Another aspect that needs to be considered is the chronological relationship between user activities. E-commerce interaction data is generally structured as a series of events that have time information, so the formation of historical features must consider the availability of information at the time of prediction. The use of activities that occur after the predicted time has the potential to cause temporal data leakage and result in overly optimistic model performance estimates. Mishra et al. (2026) discuss the importance of identifying and preventing temporal information leakage in the development of machine learning models. Although the study was conducted in different domains, the principle of temporal leakage prevention is relevant for purchase prediction research. Therefore, the formation of features based on previous history and the chronological division of data are necessary to evaluate the model's capabilities on observations from the next period.

Based on previous research, machine learning approaches have been used to predict purchase behavior and intent through various algorithms and data representations. However, these studies have differences in target definition, feature composition, and evaluation strategies. In the context of this study, there is still a need for an evaluation that specifically compares Logistic Regression and CatBoost to predict the purchase of the same SKU within a maximum of 24 hours after an add-to-cart activity, using historical and temporal behavioral features and data segmentation by time sequence. In addition, the benefits of adding user and product identities need to be evaluated to determine whether the information provides performance improvements compared to numerical features that already represent activity history.

This study aims to evaluate the utilization of historical and temporal behavioral features in predicting the likelihood of purchasing the same product within a maximum of 24 hours after add-to-cart activities. The research data comes from the Synerise RecSys Challenge 2025 which includes the activity of adding products to the cart and purchasing products. Logistic Regression was used as the baseline model and compared to CatBoost using ROC-AUC, Average Precision (AP), accuracy, precision, recall, and F1-score metrics. The research contributions include a comparison of the two algorithms on chronological data sharing, an evaluation of the addition of user identities and Stock Keeping Units (SKUs) in the CatBoost configuration, as well as feature importance analysis to identify the features that contribute the most to prediction. The results of the study are expected to provide empirical evidence regarding the use of historical and temporal behavioral features in the development of short-term purchase prediction models on e-commerce platforms.

## Methods

This study used a quantitative computational experimental approach to predict the purchase of the same SKU within 24 hours of add-to-cart activity. Two algorithms, Logistic Regression and CatBoost, were compared using historical and temporal behavioral features with time-sequence-based evaluations. The dataset came from the Synerise RecSys Challenge 2025, namely add_to_cart.parquet (7,541,117 events) and product_buy.parquet (2,318,502 events), in the period June 23–December 8, 2022. Both datasets contained client_id, sku, and timestamps. After the formation of features from the complete event history, a sample of about 2% of cart activity resulted in 149,208 modeling observations. A comparative approach to e-commerce purchase prediction was also used by Suo et al. (2026).

### 2.1. Pre-processing and Target Definition

The preprocessing stage includes checking important attributes, converting timestamps to datetime, and sequencing events by time. Binary targets are valued at 1 when there are purchases by the same user of the same SKU at intervals 0–24 hours from the basket activity, and 0 if they don't. Target matching allows for the same timestamp. The target definition is shown in Equation (1). Separation of preprocessing and evaluation processes is necessary to reduce the risk of information leakage (Apicella et al., 2025).

$$y_i = \begin{cases} 1, & if\ there\ is\ a\ purchase\ of\ the\ same\ SKU\ within\ 24\ hours\ after\ the\ activity\ i, \\ 0, & \text{if there is no such purchase.} \end{cases} \quad (1)$$

Thus, $(y_i)$ is a binary variable that indicates whether a basket activity is followed by the purchase of the same product within a specified time horizon. The target is not intended to identify all possible user purchases, but rather specifically predicts purchases related to basket activity. At the initial matching stage, the study obtained 148,875 sample activities, with 31,718 activities having matching purchases within a maximum of 24 hours. The initial matching percentage reached 21.31%. Meanwhile, the modeling dataset produced in the next stage contained 149,208 observations with a proportion of positive classes of around 21.3%.

### 2.2. Temporal Pattern Analysis and Feature Engineering

Temporal analysis is performed to understand the time lag between the activity of adding a product to the cart and purchasing the same product. For each activity that is successfully paired with a purchase, the time difference is calculated using the equation:

$$\Delta t_i = t_i^{buy} - t_i^{cart} \quad (2)$$

Behavioral features include the frequency of cart activity, the number of historical purchases of the user, the number of historical purchases of user–SKU combinations, and the recency of cart activity and purchases. The number-shaped feature is transformed by natural logarithm (Equation 3). Recency is calculated in hours, capped at 720 hours, and transformed as per Equation (4); the absence of history is valued at 720 hours before the transformation. The use of multidimensional behavior features in purchase prediction is supported by Suo et al. (2026) and Ruan and Zhai (2026).

$$F_i = ln\,(1 + C_i) \quad (3)$$

by declaring the number of cart activity before and is the value after the transformation. The $C_i F_i$ cart activity recency feature is calculated in hours and is limited to 720 hours or 30 days before being transformed as follows:

$$R_i = ln\,(1 + min(r_i, 720)) \quad (4)$$

by $r_i$ declaring the time gap since the previous cart activity. If this value is not available, a replacement value of 720 hours before the transformation is used.

The history feature is formed with an as-of join that only uses events before the predicted time; purchases on the same timestamp are not considered history. The calendar feature includes hours, days, dates, months, and weekends, with a sine-cosine cyclic representation for the time component. The limitation of the feature on the information available when the prediction is made aims to avoid temporal data leakage. The limitation of information based on time is also in line with the principle of data leakage prevention (Apicella et al., 2025).

The datasets were sorted chronologically and divided by an initial proportion of 70%/15%/15% for training, validation, and testing. 24-hour purging was applied to both subset boundaries so that the label horizon did not cross the evaluation limit. The final results were 103,553 training data, 20,719 validations, and 22,382 tests; a total of 2,554 observations were excluded from the separation interval.

### 2.3. Logistic Regression

Logistic Regression was used as a baseline with 13 numerical features. Lost values were imputed using the median of the training data, then standardized using mean and standard deviation of the training data (Equation 5). The same pre-processing statistics were applied to the validation and test data. The implementation followed the modeling practice in scikit-learn (Pedregosa et al., 2011). Logistic Regression was used as a comparative model in the purchase prediction study of Suo et al. (2026).

$$x'ij = \frac{xij - \mu_j}{\sigma_j} \tag{5}$$

The probability of purchase is estimated using a sigmoid function over a linear combination of features (Equation 6), with ß0 as the intercept and ßj as the jth feature coefficient. The model uses C = 0.1, max_iter = 3,000, and random_state = 42.

$$P(y_i = 1 \mid x_i) = \frac{1}{1+\exp(-z_i)} \tag{6}$$

with

$$z_i = \beta_0 + \sum_{j=1}^{13} \beta_j x_{ij} \tag{7}$$

In the equation, $y_i = 1$ it shows that the purchase of the same product occurred in a span of 24 hours, $x_i$ is the vector of the observation $i$-feature, $\beta_0$ is the intercept, and $\beta_j$ is the coefficient for the $j$-feature. The $z_i$ value represents the linear combination of features used to generate the prediction probability.

The classification threshold is determined from the maximum F1-score value in the validation data (Equation 7), and then applied without change to the test data.

$$t^* = \arg\max_t F_1(t) \tag{8}$$

The prediction of the class is determined using the following equation:

$$\hat{y}_i = \begin{cases} 1, & \text{if } \hat{p}_i \geq t^* \\ 0, & \text{if } \hat{p}_i < t^* \end{cases} \quad (9)$$

$\hat{p}_i$ is the predicted probability and $t^*$ is the optimal threshold obtained from the validation data. The threshold is then applied to the test data to evaluate the model's ability to identify possible purchases. The evaluation is carried out using ROC-AUC, Average Precision (AP), accuracy, precision, recall, and F1-score.

### 2.4. CatBoost

CatBoost is a decision tree-based gradient boosting algorithm designed to study nonlinear relationships and handle categorical features (Prokhorenkova et al., 2018). The study evaluated three groups of features: (1) numerical only, (2) numerical and client_id, and (3) numeric, client_id, and sku. Each group was tested with six hyperparameter combinations, resulting in 18 experiments. The range used included a depth of 4–6, a learning rate of 0.03 or 0.05, and L2 leaf regularization of 5, 7, 9, or 12 according to the combination in the notebook. The application of CatBoost for the prediction of purchasing behavior was also studied by Ruan and Zhai (2026).

Conceptually, the output of the gradient boosting model can be represented as a combination of a number of decision trees:

$$F_M(\mathbf{x}i) = \sum m = 1^M \eta h_m(\mathbf{x}_i) \quad (10)$$

by $F_M(\mathbf{x}_i)$ being the model score for observation (i), (M) is the number of trees used, $(\eta)$ is the learning rate, and $(h_m(\mathbf{x}_i))$ is the output of the decision tree (M). The score can be converted to probability using the sigmoid function:

$$P(y_i = 1 \mid x_i) = \frac{1}{1+\exp(-F_M(x_i))} \quad (11)$$

CatBoost training uses Logloss as a loss function to measure the difference between the prediction probability and the actual label. The function is formulated as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)] \quad (12)$$

where (N) is the number of observations, $(y_i)$ is the actual label, and $(P_i)$ is the model's prediction probability. The smaller the Logloss value, the better the prediction probability of representing the actual label according to the loss function. The CatBoost experiment compared three feature configurations: numerical features only, numerical features with client identities, and numerical features with client identities and SKUs.

Hyperparameter testing was performed on six combinations that included depth of 4–6, learning rates of 0.03 and 0.05, and regularization (L2) values of 5, 7, 9, and 12 according to the predetermined combinations. Each configuration used a maximum of 2,200 iterations, with an early stop after 150 iterations with no improvement in the validation evaluation metrics. The

selection of the best configuration was conducted based on the highest Average Precision (AP) value in the validation data, which was formulated as:

$$\theta^* = \arg\max_{\theta \in \Theta} AP_{\text{validation}}(\theta) \quad (13)$$

with $(\theta)$ is a set of experimental configurations and $(\theta)$ is the selected configuration. In implementation, PRAUC is used as an evaluation metric during CatBoost training, while the final selection of the configuration is performed using Average Precision calculated on the validation data. The best configuration is then used to generate predictions on the test data. As in Logistic Regression, the classification threshold is determined based on the highest F1-score on the validation data and applied to the test data. CatBoost's performance is then compared to Logistic Regression using ROC-AUC, Average Precision, accuracy, precision, recall, and F1-score. This comparison aims to find out whether the gradient boosting-based approach provides improved predictability over the baseline model.

### 2.5. Model Evaluation

The performance of both algorithms was evaluated on the same subset of tests using ROC-AUC, Average Precision (AP), accuracy, precision, recall, and F1-score. AP was prioritized for configuration selection because positive classes constituted approximately 21.3% of the modeling dataset. In unbalanced data, the Precision–Recall curve can provide more relevant information about positive class predictions than relying solely on ROC (Saito & Rehmsmeier, 2015). F1-score is calculated as the harmonic mean of precision and recall (Equation 8). Multi-metric evaluation for unbalanced classification is supported by Sujon et al. (2025).

To analyze the contribution of changes in feature composition, the AP difference is calculated using the equation:

$$F1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

The difference in AP between configurations is calculated as the second configuration AP minus the first configuration AP. The feature importance of the selected CatBoost is analyzed to describe the relative contribution of the feature, without interpreting the contribution as a cause-and-effect relationship. The results of the hyperparameter selection and the final performance value are presented in the Results and Discussion section.

## Result and Discussion

### 3.1. Characteristics and Purchasing Behavior

This experiment used two event datasets from the Synerise RecSys Challenge 2025, namely add-to-cart and product-buy. The dataset consisted of 7,541,117 product addition events to carts and 2,318,502 purchase events recorded in the period from June 23 to December 8, 2022. Each event contained a customer identity, timestamp, and SKU (Stock Keeping Unit), so that purchasing behavior could be analyzed at the customer and product level. Initial matching analysis was conducted on 148,875 samples of product addition events to the cart. The results showed that 31,718 events were followed by the purchase of the same SKU within 24 hours,

accounting for 21.31% of the overall sample. Analysis of purchase time lags showed that 48.02% of matching events were followed by purchases within five minutes, while 81.75% occurred within 30 minutes. The median purchase time lag was 5.42 minutes, while the average was 70.34 minutes. This difference shows a right-skewed distribution, meaning most purchases occur in a relatively short time, but others take much longer. The modeling dataset consists of 149,208 observations, with 31,784 positive labels and a positive class proportion of around 21.3%. The division of data by time sequence (temporal split) resulted in 103,553 observations for training data, 20,719 observations for validation data, and 22,382 observations for test data. The proportion of positive labels in each of these parts was 21.25%, 20.04%, and 22.91%. This difference in proportion shows the importance of evaluating the model using time-separated data, rather than relying solely on random data sharing (Requena et al., 2020; Abdullah-All-Tanvir et al., 2023).

### 3.2. Experimental Development and Model Configuration Selection

The dataset was divided chronologically into training, validation, and testing subsets using an initial proportion of 70%, 15%, and 15%, respectively. To minimize the risk of temporal information leakage, a 24-hour purging procedure was applied at the subset boundaries. This approach ensured that the prediction horizon did not overlap across the evaluation periods, following the principles of data leakage prevention discussed by Apicella et al. (2025). Following this procedure, the final dataset consisted of 103,553 training observations, 20,719 validation observations, and 22,382 testing observations, while 2,554 observations were excluded from the separation intervals.

The model development process involved several experimental configurations to evaluate the predictive performance of Logistic Regression and CatBoost. Logistic Regression was used as the baseline model with 13 numerical features, following its established application in machine learning classification (Pedregosa et al., 2011). Meanwhile, CatBoost was evaluated to capture potential nonlinear relationships and accommodate categorical features (Prokhorenkova et al., 2018). The use of behavioral features in purchase prediction is also consistent with previous e-commerce prediction studies (Suo et al., 2026; Ruan & Zhai, 2026).

The CatBoost experiments evaluated three feature configurations: (1) numerical features only, (2) numerical features combined with client_id, and (3) numerical features combined with client_id and sku. Each feature configuration was tested using six hyperparameter combinations involving tree depth, learning rate, and L2 leaf regularization, resulting in 18 experimental configurations.

The experiments were conducted through different development stages, including V3 and V5. Experiment V3 represented an exploratory stage in which alternative feature representations and model configurations were investigated. V5 subsequently provided the primary evaluation framework for comparing Logistic Regression and CatBoost. However, because the feature representations and evaluation procedures differed between these experimental versions, their results were not treated as a direct controlled comparison.

The best CatBoost configuration was selected based on the highest Average Precision (AP) score on the validation dataset. This metric was prioritized because the positive class represented approximately 21.3% of the modeling dataset. As discussed by Saito and Rehmsmeier (2015),

precision–recall analysis provides informative performance assessment for binary classification involving imbalanced datasets.

For both algorithms, the classification threshold was determined by maximizing the F1-score on the validation dataset and then applied unchanged to the test dataset. This procedure maintained the separation between model selection and final test evaluation.

Based on this experimental framework, V5 was designated as the primary configuration for the final model comparison. The V3 results were retained as exploratory findings, while the subsequent performance analysis focused on the V5 evaluation results.

### 3.3. Model Evaluation on V5

V5 provides a structured comparison between Logistic Regression and CatBoost. Logistic Regression was evaluated using 13 numerical features, whereas CatBoost was assessed using three feature configurations: numerical features only, numerical features combined with customer identifiers, and numerical features combined with customer identifiers and SKUs. The hyperparameter configuration was selected based on the Average Precision (AP) score obtained from the validation dataset rather than the test dataset. The best-performing CatBoost configuration used numerical features only, with a tree depth of 4, a learning rate of 0.03, and an L2 leaf regularization value of 5. This configuration achieved a validation AP of 0.3772, with early stopping identifying iteration 681 as the best iteration. The final test performance of both models is presented in Table 1. Previous studies have also compared linear and ensemble-based algorithms for consumer behavior prediction; however, performance comparisons across studies should be interpreted cautiously because of differences in datasets and target definitions (Lin, 2025; Baati & Mohsil, 2020).

Table 1. Model performance on V5 test data

**Table 1**

| Metrics | Logistic Regression | CatBoost |
|---|---|---|
| ROC-AUC | 0,6596 | 0,683 |
| Average Precision (AP) | 0,3891 | 0,4212 |
| Accuracy | 0,6665 | 0,6926 |
| Precision | 0,3539 | 0,382 |
| Recall | 0,5521 | 0,553 |
| F1-score | 0,4313 | 0,4518 |

CatBoost achieved higher values across all six evaluation metrics compared with Logistic Regression. The ROC-AUC increased by 0.0234, while AP increased by 0.0321. Accuracy improved by 0.0261, precision by 0.0281, and F1-score by 0.0205. Meanwhile, the recall improvement was relatively small, at 0.0009. These results indicate that CatBoost demonstrated better overall predictive performance, particularly in terms of discrimination

and precision–recall balance. However, its improvement in identifying positive purchase cases at the selected classification threshold was limited.

AP is particularly relevant in this evaluation because the positive class accounted for approximately 22.91% of the test dataset. Unlike accuracy, which measures the proportion of correctly classified observations at a particular threshold, AP summarizes precision across different recall levels. Therefore, AP provides a more informative assessment of positive-class ranking performance in an imbalanced classification setting. Saito and Rehmsmeier (2015) highlighted the importance of precision–recall analysis when evaluating binary classifiers on imbalanced datasets because it provides a more direct representation of positive prediction performance. Previous studies have also emphasized the importance of evaluating multiple classification metrics when predicting online shopping intentions rather than relying solely on accuracy (Abdullah-All-Tanvir et al., 2023; Baati & Mohsil, 2020).

The CatBoost AP value of 0.4212 was higher than the positive-class prevalence of 0.2291, indicating better ranking performance than a simple reference based on class prevalence. Nevertheless, this result does not establish that the model is ready for operational implementation. Further validation, probability calibration, and evaluation based on business costs and operational consequences are required before practical deployment. Previous research on boosting-based purchase prediction has also demonstrated the relevance of evaluating model performance in the context of complex consumer behavior features (Abdullah-All-Tanvir et al., 2023; Lin, 2025). However, these findings do not establish a universal advantage of boosting algorithms across different datasets and experimental settings.

The ROC curves of Logistic Regression and CatBoost on the V5 test dataset are presented in Figure 1. The ROC-AUC values of 0.6596 and 0.6830, respectively, indicate that CatBoost demonstrated slightly better discrimination between positive and negative purchase outcomes. This result suggests that CatBoost was more effective in distinguishing purchase-related cart activities from activities that were not followed by a purchase within the defined 24-hour period. The use of temporally separated evaluation data is also relevant in clickstream-based and time-sequence prediction research, where model performance needs to be assessed on observations separate from the training data (Requena et al., 2020; Zhou & Hudin, 2024).

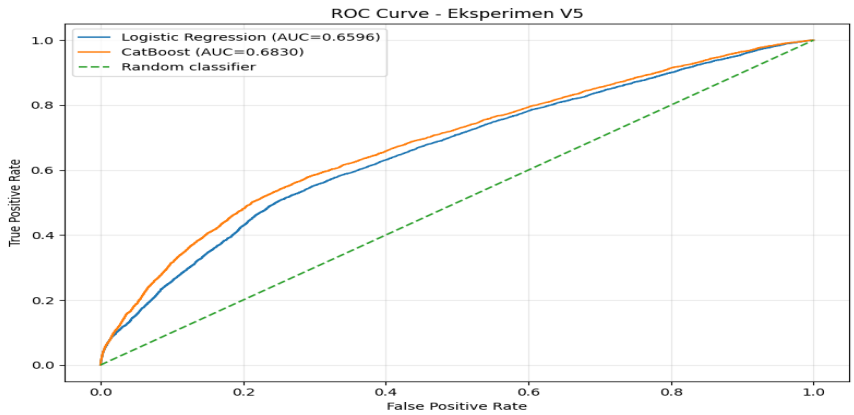



Figure 1. ROC Curve Experiments

The Precision–Recall curves of both models are presented in Figure 2. CatBoost achieved an AP of 0.4212, compared with 0.3891 for Logistic Regression. Both models obtained AP values above the positive-class prevalence of 0.2291, indicating better positive-class ranking performance than a simple prevalence-based reference. CatBoost also achieved a higher AP value, suggesting that it provided a better precision–recall trade-off across different classification thresholds. This finding supports the use of Precision–Recall analysis as an important evaluation tool for imbalanced classification problems (Saito & Rehmsmeier, 2015).

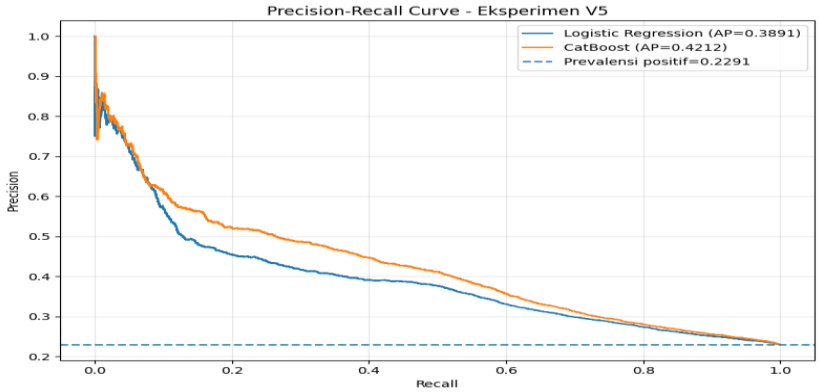



Figure 2. Precision-Recall Curve Experiments

The results of V5 evaluation indicates that CatBoost outperformed Logistic Regression across all six reported test metrics. The improvement was particularly evident in ROC-AUC and Average Precision, while the recall difference remained relatively small. These findings suggest that CatBoost provided better overall predictive performance for identifying purchase-related cart activities within the defined 24-hour prediction horizon. However, the results should be interpreted within the scope of the dataset, feature configurations, and evaluation procedures used in this study. Further experiments involving feature ablation, probability calibration, and operational validation are necessary to assess the model's generalizability and practical applicability.

### 3.4. Comparison of Feature Sets and Feature Importance Analysis

To understand the factors that contribute the most to CatBoost prediction, feature importance analysis was performed on the selected model. The results are presented in Table 2.

Table 2. Features with the Greatest Contribution to the CatBoost V5 Model

**Table 2**

| Features | Feature Importance (%) |
|---|---|
| Historical user purchase count | 39,38% |
| Frequency of cart activity | 20,50% |
| User purchase recency | 13,00% |
| Recency cart activity | 7,52% |
| The same historical number of SKU purchases by users | 4,76% |
| Recency of the purchase of the same SKU | 3,57% |

| | |
|---|---|
| Hours of activity (cosine component) | 2,83% |
| Month of activity | 2,34% |
| Dates in months | 1,94% |

The results showed that the number of users' historical purchases had the largest contribution with a feature importance value of 39.38%. This feature was followed by the frequency of cart activity of 20.50% and the user's purchase recency of 13.00%. These three features together represent more than 70% of the total feature importance model. The dominance of user purchase history features shows that previous purchase behavior is important information in predicting possible purchases after the addition of products to the cart. Users with certain historical purchase patterns may have different characteristics in completing shopping activities. Purchase prediction research using XGBoost as well as explainable machine learning studies also use feature importance to identify informative behavioral signals (Wang et al., 2023; Chen et al., 2023; Lin, 2025).

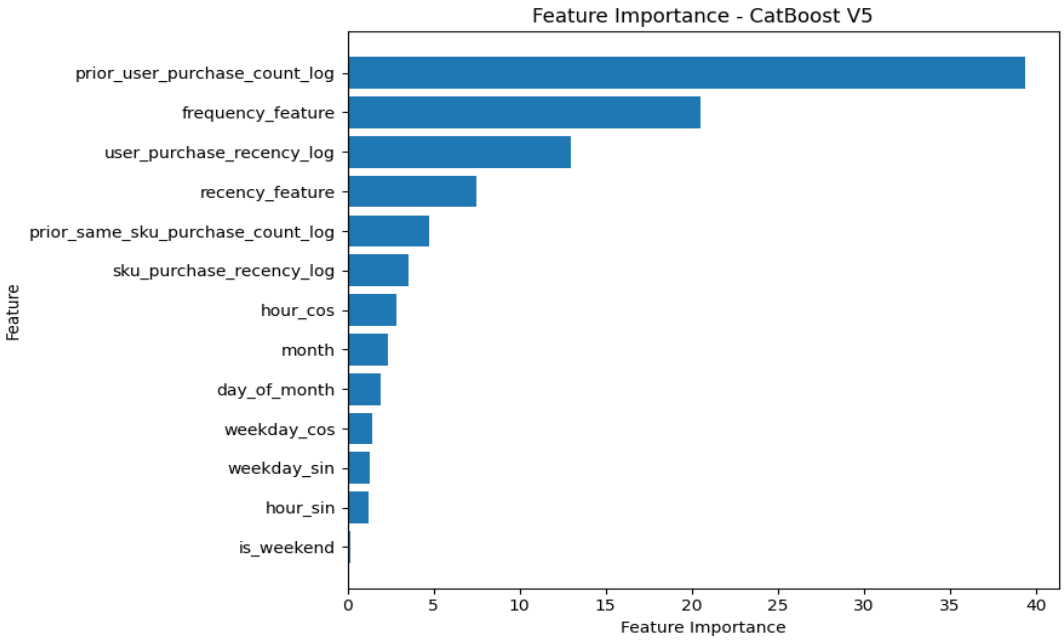



Figure 3. Feature Importance CatBoost V5 Diagram

Meanwhile, the frequency of cart activity and the recency of the cart provide information about the intensity and timing of user activity. The same SKU purchase history feature also contributes to the prediction, although the contribution is smaller than the overall user purchase history. This shows that the purchase history at the user level and at the user-product combination level provide different information for the model. Compared to historical behavior features, calendar features such as hours, months, dates, and days of the week have a smaller contribution. For example, the clock feature in *cosine* form has a feature importance value of 2.83%, while the weekend indicator is only 0.20%. These findings indicate that, in the model configuration tested, historical behavior information is more dominant than most calendar information in the prediction process. However, feature importance does not indicate a cause-and-effect relationship and cannot be directly used to conclude that a change in a feature will

increase or decrease the likelihood of a purchase. The value describes the relative contribution of features to the trained CatBoost model, rather than causal influences on consumer behavior.

The findings of the study show that the CatBoost-based approach provides better purchase prediction results than Logistic Regression on the test data. These advantages are consistent in ROC-AUC, AP, accuracy, precision, recall, and F1-score. These results indicate that tree-based models are able to utilize patterns in historical and temporal behavioral features more effectively than linear models in the experiments conducted. One possible explanation is CatBoost's ability to capture non-linear relationships and inter-feature interactions. The relationship between cart activity frequency, user purchase history, and activity time proximity does not have to be linear. CatBoost can learn these patterns through the formation of a decision tree, whereas Logistic Regression models the relationship between features and log-odds through a linear combination of features used.

In addition, the results of the ablation analysis showed that the addition of user identities and SKUs did not increase validation APs compared to numerical configurations alone. These findings are important because they show that high-resolution identity information does not necessarily provide additional benefits in temporal prediction. One possible explanation is the large number of user and product identities with limited observation history. However, this explanation is still an interpretation that requires further testing, rather than a direct conclusion from the ablation results. The contribution of the feature also shows that the user's purchase history is more dominant than the calendar feature. Thus, the results of the study strongly support the importance of historical behavioral information rather than the claim that the calendar time feature is the primary determinant of purchases. Temporal features remain relevant as part of the representation of activity, but their additional benefits need to be evaluated in isolation through special ablation experiments. The results also need to be interpreted by considering the differences in the prevalence of positive classes between subsets. The proportion of positive classes in the test data was 22.91%, higher than the validation data of 20.04%. These distribution changes indicate variations in data characteristics between periods that can affect prediction performance. Although temporal separation provides a more realistic evaluation than random distribution, testing on other periods or datasets is still required to assess model stability. XGBoost-based purchase prediction studies place feature engineering and selection as an important part of modeling; the benefits of additional features still need to be tested on consistent data sharing (Wang et al., 2023).

### 3.5. Implications and Limitations of Research

The results show that purchase predictions over a 24-hour time span can benefit from nonlinear modeling of customer behavioral features. In V5, CatBoost outperformed Logistic Regression based on reported test metrics. Meanwhile, a comparison of feature sets showed that the addition of customer identities and SKUs did not increase validation APs compared to the use of numerical features alone. The findings support the use of structured behavioral features as a practical starting point in the development of purchase prediction models.

However, this study still has some limitations. Target labels are determined based on matching product addition to cart events with the same SKU purchases by the same customer within a 24-hour time frame. Product addition to cart events and repeated purchases can create

matching ambiguity if the research procedure does not explicitly determine how each purchase is paired with the product add to cart event. The initial match analysis and modeling dataset have different numbers of positive labels. These differences need to be explained based on sampling procedures and dataset formation before the article is published. The comparison of V3 and V5 has not been fully controlled because the results of the Logistic Regression baseline model reported are different. Finally, current experiments have not separated the influence of calendar features from customer behavior history and have not proven the model's generalizability on other datasets or periods. Online purchase prediction studies also emphasize the need to assess the data context and target definition before predictive results are used operationally (Lin, 2025; Baati & Mohsil, 2020).

Further research needs to prioritize controlled ablation studies by maintaining consistent data sharing, pre-processing, classification threshold determination strategies, and model tuning procedures. Data matching procedures also need to be validated to reduce the likelihood of ambiguous label formation. In addition, the final model needs to be evaluated over additional temporal periods. These measures are expected to provide stronger evidence regarding the contribution of behavioral and temporal features to purchase predictions in the context of e-commerce.

## Conclusion

This study evaluated the use of behavioral and temporal features to predict product purchases on e-commerce platforms within 24 hours after add-to-cart by comparing Logistic Regression and CatBoost algorithms. The experiment was conducted using user activity data from the Synerise RecSys Challenge 2025 with data segmentation based on time order to evaluate the model's performance on observations from the following period. The results of the V5 experiment showed that CatBoost performed better than Logistic Regression in all six test metrics used. CatBoost obtained a ROC-AUC value of 0.6830, Average Precision (AP) of 0.4212, accuracy of 0.6926, precision of 0.3820, recall of 0.5530, and F1-score of 0.4518. Compared to Logistic Regression, CatBoost increased the AP value by 0.0322 and the ROC-AUC by 0.0234. These results showed that CatBoost had better predictive ranking and class discrimination capabilities on the evaluated test data. Feature importance analysis showed that the user's historical purchase number, cart activity frequency, and time interval since the user's last purchase were the features that contributed the most to prediction in the selected CatBoost model.

These findings suggest that the user's historical behavioral information provides an important signal for predicting possible purchases after the product is added to the cart. In contrast, calendar-based features have a relatively smaller contribution. In addition, the addition of customer identities and SKUs did not increase the validation AP value compared to configurations that used only numerical features. This suggests that such additional identity information has not provided measurable benefits in the evaluated experimental configuration. Nonetheless, the results need to be interpreted taking into account the limitations of the experiment. The difference in the number of positive labels between the initial matching results and the modeling dataset still needs further explanation. In addition, comparisons between previous experimental versions have not been fully controlled, and this study has not isolated the contribution of each group of behavioral features and calendars through a thorough

ablation analysis. Further research needs to validate event matching procedures, conduct feature ablation experiments under consistent conditions, and evaluate model performance over temporal periods and additional datasets. Evaluation of probability calibration as well as operational costs and consequences is also required before the model is considered for practical application. Overall, the study shows that CatBoost is a promising approach to predicting short-term purchasing behavior based on a history of e-commerce activity, although further testing is still needed to confirm its generalization capabilities and practical benefits.

## Acknowledgments

The author would like to thank the Data Science Study Program, State University of Surabaya, for the academic environment that supports the implementation of this research. The author also expresses his appreciation to the organizers of the Synerise RecSys Challenge 2025 for providing the e-commerce activity dataset used in this study. The dataset allows the author to examine the use of behavioral and temporal features in short-term purchase predictions.

## References

Synerise. (2025). Synerise *RecSys Challenge 2025 dataset* [Data set]. https://recsys.synerise.com/data-set

Alamuri, O. R. S. G., & Bondalapu, C. K. (2026). Predicting e-commerce purchase intention using machine learning. *ASEAN Journal of Scientific and Technological Reports, 29*(1). https://doi.org/10.55164/ajstr.v29i1.260159

Bastos, J. A., & Bernardes, M. I. (2024). Understanding online purchases with explainable machine learning. Information, *15*(10), Article 587. https://doi.org/10.3390/info15100587

Chopra, I. P., Jebarajakirthy, C., Jain, T., & Maseeh, H. I. (2024). Electronic shopping cart abandonment: What do we know and where should we be heading? *Electronic Markets, 34*, Article 25. https://doi.org/10.1007/s12525-024-00697-0

Kafilla, P. I., Utomo, F. S., & Karyono, G. (2025). Enhancing customer purchase behavior prediction using PSO-tuned ensemble machine learning models. *Journal of Informatics Engineering (JUTIF), 6*(5), 3509–3525. https://doi.org/10.52436/1.jutif.2025.6.5.4952

Mishra, L. N., Rangari, A., Nagrare, S., & Nayak, S. K. (2026). A taxonomy for detecting and preventing temporal data leakage in machine learning-based build prediction: A dual-platform empirical validation. *PLOS ONE, 21*(5), Article e0340167. https://doi.org/10.1371/journal.pone.0340167

Wen, Z., Lin, W., & Liu, H. (2023). Machine-learning-based approach for anonymous online customer purchase intentions using clickstream data. *Systems, 11*(5), Article 255. https://doi.org/10.3390/systems11050255

Apicella, A., Isgrò, F., & Prevete, R. (2025). Don't push the button! Exploring data leakage risks in machine learning and transfer learning. Artificial Intelligence Review, 58, Article 339. https://doi.org/10.1007/s10462-025-11326-3

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In Advances in Neural Information Processing Systems (Vol. 31). https://proceedings.neurips.cc/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html

Ruan, W., & Zhai, Z. (2026). Predicting user purchase behavior via multidimensional feature engineering and explainable machine learning. Journal of Theoretical and Applied Electronic Commerce Research, 21(10), Article 352. https://doi.org/10.3390/jtaer21100352

Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A. (2025). Accuracy, precision, recall, f1-score, or MCC? Empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. Journal of Big Data, 12, Article 268. https://doi.org/10.1186/s40537-025-01313-4

Suo, L., Xia, L., Chung, Y., & Kim, E. (2026). Layered feature engineering for e-commerce purchase prediction: A hierarchical evaluation on Taobao user behavior datasets. Computers, Materials & Continua, 87(1), Article 78. https://doi.org/10.32604/cmc.2025.076329

Abdullah-All-Tanvir, Khandokar, I. A., Islam, A. K. M. M., Islam, S., & Shatabda, S. (2023). A gradient boosting classifier for purchase intention prediction of online shoppers. Heliyon, 9(4), e15163. https://doi.org/10.1016/j.heliyon.2023.e15163

Zhou, S., & Hudin, N. S. (2024). Advancing e-commerce user purchase prediction: Integration of time-series attention with event-based timestamp encoding and graph neural network-enhanced user profiling. PLOS ONE, 19(4), e0299087. https://doi.org/10.1371/journal.pone.0299087

Baati, K., & Mohsil, M. (2020). Real-time prediction of online shoppers' purchasing intention using random forest. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial intelligence applications and innovations* (Vol. 583, pp. 43–51). Springer. https://doi.org/10.1007/978-3-030-49161-1_4

Chen, Y., Liu, H., Wen, Z., & Lin, W. (2023). How explainable machine learning enhances intelligence in explaining consumer purchase behavior: A random forest model with anchoring effects. *Systems, 11*(6), Article 312. https://doi.org/10.3390/systems11060312

Lin, J. (2025). Application of machine learning in predicting consumer behavior and precision marketing. *PLOS ONE, 20*(5), Article e0321854. https://doi.org/10.1371/journal.pone.0321854

Requena, B., Cassani, G., Tagliabue, J., Greco, C., & Lacasa, L. (2020). Shopper intent prediction from clickstream e-commerce data with minimal browsing information. *Scientific Reports, 10*, Article 16983. https://doi.org/10.1038/s41598-020-73622-y

Wang, W., Xiong, W., Wang, J., Tao, L., Li, S., Yi, Y., Zou, X., & Li, C. (2023). A user purchase behavior prediction method based on XGBoost. *Electronics, 12*(9), Article 2047. https://doi.org/10.3390/electronics12092047

Yang, Y. (2025). Data-driven prediction of future purchase behavior in cross-border e-commerce using sequence modeling with PSO-tuned LSTM. *PLOS ONE, 20*(12), Article e0337932. https://doi.org/10.1371/journal.pone.0337932