ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

An Enhanced LightGBM Framework for E-commerce Purchase Prediction

Zhiyuan Zhu[1,a,#], Kunlin Li[2,b,#], Ziteng Wang[3,c,#], Xinwen Lyu[4,d,#]

[1]*Edmonds College, Edmonds, 98036, Washington, United States*
[2]*Universiti Sains Malaysia, Gelugor, Penang, 11700, Malaysia*
[3]*Shanghai International Studies University, Shanghai, 200083, China*
[4]*Wuxi Dipont School of Arts and Science, Wuxi, Jiangsu, China*
[a]*zhuzhi45850@student.edmonds.edu,* [b]*kunlin2026@student.usm.my,* [c]*3396559839@qq.com,*
[d]*18552137076@163.com*
[#]*Co-first author*

Abstract: *E-commerce purchase prediction is often evaluated as a classification problem, but in many business settings, the predicted probability is also used to rank customers, allocate budgets, and estimate expected returns. This study develops an Enhanced LightGBM framework to address three practical problems: limited representation of behavioral features, limited use of categorical information, and unreliable probability outputs. The framework combines 15 domain-informed derived features, leakage-free five-fold target encoding, probability averaging across multiple random seeds, isotonic probability calibration, and threshold re-optimization. Experiments were conducted on a publicly available e-commerce dataset with 100,000 records. The results show that different modules contribute in various ways. Feature expansion and target encoding provide only a small gain in discrimination, with AUC increasing from 0.7603 to 0.7636. Probability calibration has the clearest effect: the Brier Score decreases from 0.2165 to 0.1922, an improvement of 11.2%. Threshold optimization moves the operating point toward higher recall, increasing recall from 58.39% to 82.64% and producing an F1 score of 0.6801, while precision decreases to 57.77%. These results suggest that the main value of the proposed framework is not a large improvement in ranking ability. Instead, it improves the reliability of probabilities and model reproducibility, which can better support marketing decisions based on expected value.*

Keywords: *e-commerce purchase prediction; LightGBM; feature engineering; target encoding; probability calibration; threshold selection*

1. Introduction

As e-commerce platforms continue to expand, predicting whether a user will make a purchase has become an important task in recommendation systems, targeted marketing, and resource allocation [1]. If a platform can estimate purchase intention before the final decision, it can better identify potential customers and manage advertising or promotion budgets. However, this task is different from a simple classification problem. In many real applications, the model output is not only used to predict whether a customer will purchase. Purchase probabilities can also support customer prioritization, personalized marketing, and other cost-sensitive business decisions [1,2]. Therefore, the reliability of the predicted probability is important in addition to classification accuracy.

Purchase behavior is influenced by many factors, including user characteristics, price and product information, and browsing patterns. These variables can exhibit nonlinear relationships and strong interactions, making purchase prediction a challenging tabular data problem. Tree-based gradient boosting methods are widely used for this type of data because they usually perform strongly on structured datasets [3]. LightGBM is especially useful because its histogram-based learning, Gradient-based One-Side Sampling, and Exclusive Feature Bundling improve training efficiency while keeping competitive predictive performance [4].

However, directly applying LightGBM has several practical limitations. First, many raw behavioral variables in e-commerce datasets are recorded as simple action counts, such as page views, favorites, and add-to-cart behaviors [5]. Each variable reflects only one aspect of user interest, so the model must learn their combinations. Second, e-commerce data commonly contain categorical variables. One-hot encoding

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

shows category differences, but it does not directly represent different base purchase rates across categories. Target encoding can provide this information, but careless implementation can lead to target leakage and overly optimistic evaluation [6,7]. Earlier work on high-cardinality categorical features also shows why regularized target-based summaries can be useful when categories are numerous [8]. CatBoost further demonstrates that target statistics must be constructed carefully to reduce prediction shift and target leakage [9].

Third, LightGBM training includes random procedures such as feature and row subsampling. Different random seeds can therefore produce different tree structures and early stopping points [10]. Also, class weighting can improve recognition of positive cases but may distort the scale of output scores, so a score such as 0.8 may not represent an actual 80% purchase probability [2]. Previous calibration studies have shown that classifier scores often need an additional calibration step before they can be interpreted as probabilities [11,12].

To address these issues, this paper develops an Enhanced LightGBM framework with improvements at three levels. At the representation level, 15 domain-informed features are constructed, and a leakage-free five-fold target encoding procedure is used. At the training level, the model is trained with several random seeds and the predicted probabilities are averaged to reduce random fluctuation. At the output level, isotonic regression is used for probability calibration, followed by a new search for the classification threshold. The framework does not introduce a new gradient boosting algorithm. Its contribution is to organize existing methods into a complete and reproducible workflow and to use ablation experiments to clarify the role of each component.

The experimental results also show an important distinction. The large increase in recall is mainly due to class weighting and threshold adjustment, not to a major improvement in the model's ranking performance. In contrast, probability calibration gives a clear improvement in Brier Score. Based on these findings, this paper positions the framework primarily as a method for improving probability reliability and reproducibility rather than as one that substantially increases classification discrimination.

2. Method

2.1 Overall Framework

The Enhanced LightGBM framework contains three connected stages. The first stage improves feature representation through data cleaning, leakage control, 15 domain-informed derived features, and leakage-free five-fold target encoding [6]. The second stage improves training stability by using class weights, hyperparameter optimization, multiple random seeds, and probability averaging [10]. The third stage improves output quality through isotonic regression calibration [2] and a new threshold search on the calibrated probabilities. The full workflow is shown in Figure 1.

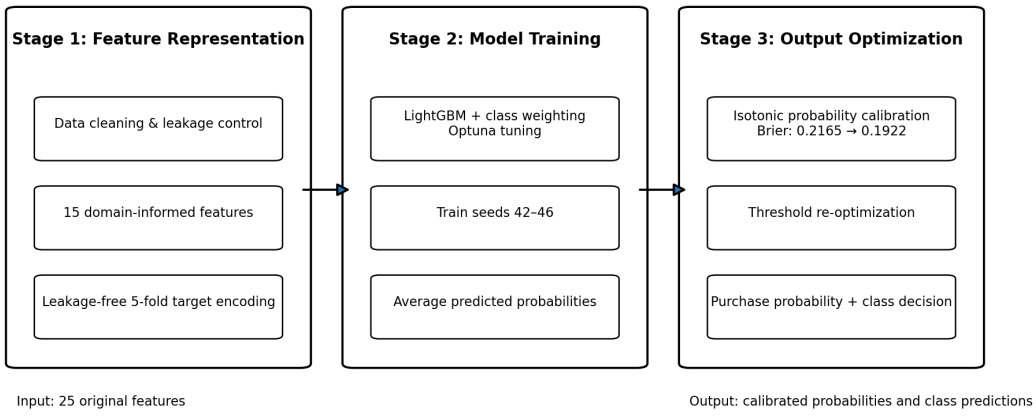


Figure 1. Overall workflow of the Enhanced LightGBM framework

The order of these stages is important. Target encoding must be created after the data is split because the category statistics should come only from the training data. Probability calibration is placed after multi-seed averaging because the calibration model should receive a more stable probability estimate. Finally, threshold search must be completed after calibration because isotonic regression is monotonic but nonlinear. It changes the shape and scale of the probability distribution, so a threshold selected before

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

calibration may no longer be appropriate.

2.2 Domain-Informed 15-Feature Expansion

Most original behavioral variables are counts of a single action. Used alone, they do not fully describe the process from viewing a product to forming purchase intention. Therefore, 15 derived features were grouped into six categories: interaction intensity, price perception, purchase intention, browsing quality, social influence, and product attractiveness. Three principles were followed. First, every derived feature uses only information observable before the purchase decision, so the purchase label is never used. Second, log transformations are applied to long-tailed count variables to reduce the influence of extreme values. Third, ratio features are created when possible so users with different overall activity levels can be compared more fairly (Table 1).

Table 1. Domain-Informed Derived Features and Their Definitions

Category	**Feature**	**Definition / Rule**	**Purpose**
Interaction	Total Interaction	Like + Comment + Collect + Share	Overall interest intensity
Interaction	Log Interaction	log(1 + Total Interaction)	Reduce long-tail influence
Interaction	Engagement Rate	Total Interaction / View Number	Interest per exposure
Interaction	Log Like Number	log1p(Like Number)	Reduce skew in likes
Price	Discount Amount	Original Price − Current Price	Absolute discount
Price	Discount Ratio	(Original Price − Current Price) / Original Price	Relative discount
Price	Price Mid-range	1 if 200 ≤ price ≤ 500, else 0	Mark a moderate price band
Price	Price After Discount	Price × (1 − discount rate)	Estimated actual price
Intent	Intent Score	Add-to-cart×3 + coupon received×2 + follow author×1.5	Combined purchase intention
Intent	High Value User	1 if User Level ≥ 6	Identify higher-value users
Browsing	Browse Efficiency	Purchase-related actions / View Number	Purposeful browsing
Browsing	Hesitation Index	PV count×(1−add-to-cart)×(Total Interaction==0)	Repeated browsing without action
Social	Social Influence	Fans Number − Follow Number	Net social influence
Product	Collect per View	Collect Number / View Number	Collection ability per exposure
Product	Content Richness	Image count + has-video×2 + title length/10	Amount of product information

Ratio features are especially useful for separating absolute activity from relative interest. For example, a high total interaction count can mean that a user strongly likes a product, but it can also mean that the user is generally very active. Dividing total interactions by page views yields an engagement rate that better reflects the interest generated by each exposure. Browse efficiency and the hesitation index use the same logic. Log transformations are used because a small number of highly active users can have interaction counts much larger than the majority of users, which can otherwise make tree splits focus too much on the tail of the distribution.

All 15 features are computed only from original input variables and do not use the purchase label. Therefore, the feature expansion itself does not create target leakage. After expansion, the number of input features increases from 25 to 40.

2.3 LightGBM Base Model and Hyperparameter Optimization

LightGBM is selected as the base model because it performs well with structured tabular data [3], captures nonlinear interactions, and remains computationally efficient after feature expansion [4]. To increase sensitivity to positive purchase cases, class weights are included during training. Hyperparameters are optimized with Optuna [13], including the number of trees, learning rate, maximum depth, number of leaves, regularization parameters, and class weight. The validation F1 score is used as the optimization target. Because class weighting changes the training objective, the raw output score may

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

become less aligned with the true purchase probability [2,12]. This is one reason probability calibration is included later in the framework.

2.4 Leakage-Free Five-Fold Target Encoding

Categorical variables often contain useful business information because different categories can have different base purchase rates. Target encoding replaces each category with a smoothed estimate of its historical target rate [6,8]. In this study, the smoothed target encoding for category c is defined as:

$$TE(c) = \frac{(n_c \cdot y_c + \alpha \cdot P)}{(n_c + \alpha)} \tag{1}$$

where n_c is the number of training samples in categor y_c, $\bar{y}_c$ is the mean purchase rate for that category, P is the overall purchase rate in the training set, and α is the smoothing parameter. When a category has many samples, the encoded value is close to its own observed purchase rate. When the category is small, the value is pulled toward the global purchase rate, which can reduce overfitting.

The main risk of target encoding is leakage. If the encoding of a training observation is computed using the full training set, that observation's label contributes to its encoded value. The model then receives information that would not be available for a new observation [6,7]. To avoid this problem, five-fold out-of-fold encoding is used. For each fold, category statistics are calculated from the other four folds and then applied to the held-out fold. In this way, the encoded value for each training observation is generated without using its own target label. For the validation and test sets, category mappings are calculated only from the full training set. This procedure follows the general principle that target-derived features for a training observation should not be constructed using that observation’ s own label [9].

After target encoding, the original categorical variable is replaced. The final input, therefore, remains 40-dimensional: 24 original non-categorical variables, 15 derived variables, and 1 target-encoded categorical feature.

2.5 Multi-Seed Probability Averaging

LightGBM can give slightly different results across random seeds because feature subsampling, row subsampling, and early stopping include random variation [10]. This variation may appear small in a single experiment, but it matters when a model is retrained regularly or when reported results should be reproducible.

To reduce this variation, five models are trained with random seeds 42, 43, 44, 45, and 46. All other settings, including the data split, features, and hyperparameters, are kept the same. The final probability is the arithmetic mean of the five outputs:

$$\bar{p} = \frac{1}{K}\sum_{k=1}^{K} p_k \tag{2}$$

Probability averaging is used instead of majority voting because it preserves continuous confidence information. This is necessary for the later calibration step. The main cost is additional training time and storage, so this module exchanges computational efficiency for better stability.

2.6 Probability Calibration and Threshold Re-optimization

In marketing applications, a prediction score can be used not only to classify customers but also to estimate expected return or assign a marketing budget. In these settings, the numerical value of probability is meaningful [2,11]. However, class weighting can shift the distribution of model scores, so a raw score of 0.8 does not necessarily mean that about 80% of similar users will purchase.

This study uses isotonic regression for probability calibration [2]. Isotonic regression learns a monotonic piecewise-constant mapping from raw scores to calibrated probabilities by minimizing squared error on the validation data. Compared with methods that assume a sigmoid mapping, isotonic regression makes fewer assumptions about the calibration curve and fits more complex distortions when enough calibration data are available [11,12]. Probability quality is evaluated using the Brier Score [14]:

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{3}$$

where p_i is the predicted probability and y_i is the true binary label. A lower Brier Score means the probability estimates are closer to the observed outcomes.

Because isotonic regression is monotonic, it generally keeps the ordering of observations and therefore should not materially change rank-based discrimination. However, it changes the probability scale, so the previous 0.5 threshold may no longer represent a useful operating point. The threshold is therefore searched again on the calibrated validation probabilities to maximize F1:

$$\theta^* = argmax_\theta\, F1(\theta),\ \hat{y} = 1[\bar{p}_cal \geq \theta^*] \tag{4}$$

This design separates probability estimation from the final decision rule. Calibration makes probability values more interpretable, while threshold selection chooses an operating point that matches the objective. The two steps should be treated as connected but different tasks.

3. Experimental Design

3.1 Dataset

The experiments use a public social e-commerce user behavior dataset from Alibaba Cloud Tianchi [15]. The data include user profiles, product information, and pre-purchase behaviors such as browsing, interaction, and collection. Purchase prediction is modeled as a binary classification task, where 1 denotes a final purchase, and 0 denotes no purchase. After filtering, the dataset contains 100,000 observations: 44,983 purchase cases and 55,017 non-purchase cases. The positive-class proportion is about 44.98%, so the dataset is not severely imbalanced. After the initial variable screening, 25 original input features remain.

3.2 Variable Screening and Leakage Control

A purchase prediction model should use only information that is available before the purchase decision. Variables created after the decision can cause data leakage and make offline evaluation too optimistic [7]. For this reason, six variables are removed or excluded for different reasons, as shown in Table 2.

Table 2. Variable Screening and Leakage-Control Decisions

Variable	Action	Reason
label	Used only as target	Directly represents the purchase outcome
user_id	Removed	High-cardinality identifier; may encourage memorization of users
item_id	Removed	High-cardinality identifier; may encourage memorization of items
purchase_intent	Removed	Can be reconstructed from other behavioral variables; redundant
interaction_rate	Removed	Highly related to interaction variables; redundant for interpretation
coupon_used	Removed	No full timestamp; could occur at purchase time and create leakage

These variables are not all removed because of target leakage. User ID and item ID do not directly contain the purchase result, but a tree model may memorize specific users or products and then generalize poorly to unseen cases. Purchase intent and interaction rate are removed because their information can be reconstructed from other features. The coupon used is different: because the dataset does not provide complete timestamps, it is not possible to confirm whether coupon use occurred before the purchase decision. It is therefore removed as a precaution against true leakage.

3.3 Data Split

The data are divided into training, validation, and test sets. First, 80% of the full dataset is assigned

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

to the development portion and 20% to the test set. The development portion is then divided 80:20 into training and validation sets. As a result, the final proportions are 64% training, 16% validation, and 20% test. The training set is used to fit the model, the validation set is used for hyperparameter selection, probability calibration, and threshold selection, and the test set is used only for final evaluation [7,10].

3.4 Baseline Models

Eight common machine learning algorithms are used as baselines: logistic regression, k-nearest neighbors, decision tree, random forest, gradient boosting decision tree (GBDT), AdaBoost, Gaussian Naive Bayes, and linear support vector machine. These models cover linear, distance-based, single-tree, ensemble-tree, boosting, and probabilistic approaches. All baselines use the same data split and input feature set to ensure a more consistent comparison.

3.5 Comparison Settings

Three model levels are compared. The first level contains the eight baseline models. The second level is Improved LightGBM, which adds the 15 derived features, class weighting, Optuna hyperparameter optimization, and threshold adjustment. The third level is Enhanced LightGBM, which further adds five-fold target encoding, multi-seed probability averaging, and isotonic probability calibration. To examine which component is responsible for each result, an ablation study is also conducted with five configurations: A is the Improved LightGBM base; B adds target encoding; C adds multi-seed ensembling; D adds probability calibration; and E adds threshold re-optimization.

3.6 Evaluation Metrics

Several complementary metrics are used. Accuracy measures the overall percentage of correct predictions, but it may hide differences in positive-class recognition. Precision measures how many predicted purchasers are true purchasers, so lower precision can mean more wasted marketing contacts. Recall measures how many true purchasers are successfully identified. In a marketing setting, recall can be important when a likely customer is missed, as there is a meaningful opportunity cost. F1 summarizes the precision-recall tradeoff. Precision-recall analysis is particularly useful when the positive class is the main operational focus [16].

Two threshold-independent metrics are also reported. AUC measures the model’ s ability to rank positive cases above negative cases and is not changed simply by selecting a different classification threshold. The Brier Score measures squared error between predicted probabilities and actual labels [14]. Reporting AUC, F1, and Brier Score together helps separate three different questions: ranking ability, classification operating point, and probability quality.

4. Results and Analysis

4.1 Baseline Results

The test-set results of the eight baseline models are shown in Table 3. Accuracy is generally between 0.69 and 0.71. GBDT gives the highest F1 score among the baselines at 0.6389, with an AUC of 0.7603. However, its recall is only 58.39%, meaning that more than 40% of actual purchasers are not identified. For a marketing task that values broader coverage of potential customers, this recall level may be too low.

Table 3. Performance of Baseline Models on the Test Set

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.6963	0.7070	0.5549	0.6217	0.7466
KNN	0.6789	0.6830	0.5342	0.5995	0.7152
Decision Tree	0.6917	0.6903	0.5704	0.6247	0.7371
Random Forest	0.7027	0.7110	0.5711	0.6334	0.7580
GBDT	0.7032	0.7055	0.5839	0.6389	0.7603
AdaBoost	0.7038	0.7085	0.5804	0.6381	0.7605
Gaussian NB	0.6857	0.7450	0.4582	0.5674	0.7331
Linear SVM	0.6970	0.7156	0.5417	0.6166	0.7467

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

4.2 Improved and Enhanced LightGBM

Table 4 compares the best baseline with the two LightGBM versions. Improved LightGBM and Enhanced LightGBM both reach an F1 score of 0.6801, compared with 0.6389 for GBDT. Recall increases from 58.39% to above 82%.

Table 4. Performance Comparison of GBDT, Improved LightGBM, and Enhanced LightGBM

Model	Accuracy	Precision	Recall	F1	AUC
GBDT (best baseline)	0.7032	0.7055	0.5839	0.6389	0.7603
Improved LightGBM	0.6488	0.5761	0.8301	0.6801	0.7633
Enhanced LightGBM	0.6502	0.5777	0.8264	0.6801	0.7636

Figure 2 shows the main classification metrics across selected models.

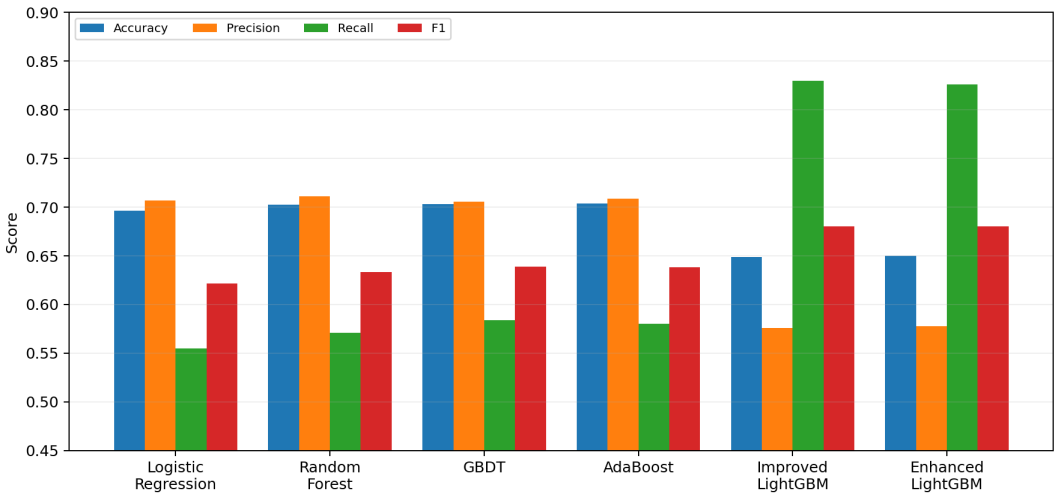


Figure 2. Main classification metrics across selected models

If only F1 and recall are considered, this improvement may look like a large gain in predictive ability. However, the other metrics give an important constraint. Accuracy decreases from 0.7032 to 0.6502, precision decreases from 70.55% to 57.77%, and AUC rises by only 0.0033, from 0.7603 to 0.7636. Because AUC reflects ranking ability rather than a chosen classification threshold, this small AUC change suggests that the model did not learn much additional discriminative information.

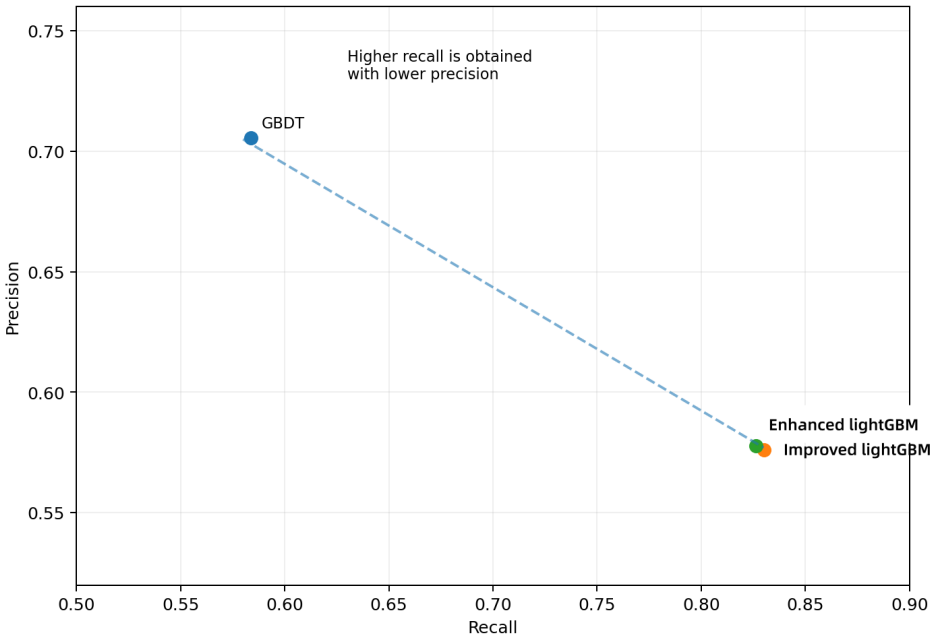


Figure 3. Operating points in the precision-recall plane

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

The stronger explanation is that class weighting and threshold adjustment move the decision boundary toward the positive class. More users are classified as likely purchasers, increasing recall but also producing more false positives and lowering precision. This is a change in the operating point on the precision-recall trade-off rather than a large shift in the model's overall ranking ability.

The operating points of Improved LightGBM and Enhanced LightGBM are almost the same, and their F1 scores are exactly equal. This means the target encoding, multi-seed averaging, and probability calibration modules do not create a meaningful classification. Their contribution should therefore be evaluated alongside other outcomes, such as probability, quality, and stability. (Figure 3)

4.3 Ablation Study

The ablation results are shown in Table 5 and Figure 4. Across configurations A to E, F1 changes only from about 0.680 to 0.681. In contrast, the Brier Score decreases from 0.218 to 0.192. This difference helps separate the function of each module.

Table 5. Ablation Results for Enhanced LightGBM Modules

Configuration	Added Module	F1	Brier Score
A	Improved LightGBM base	0.680	0.218
B	+ Target Encoding	0.680	0.217
C	+ Multi-seed Ensemble	0.681	0.217
D	+ Probability Calibration	0.626	0.192
E	+ Threshold Optimization	0.681	0.192

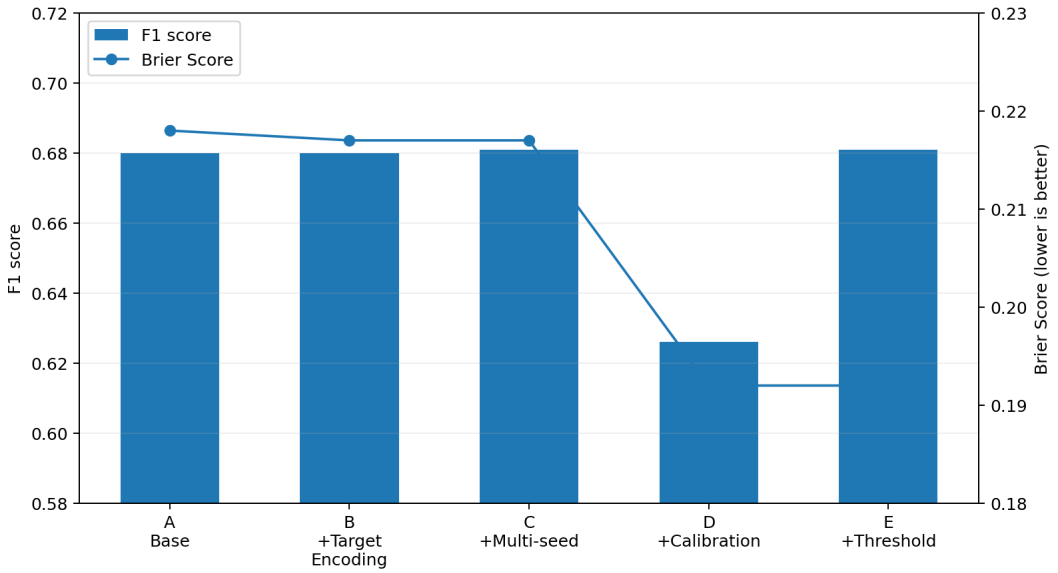


Figure 4. F1 and Brier Score changes in the ablation study

Target encoding (configuration B) leaves F1 at 0.680 and reduces Brier Score only slightly, from 0.218 to 0.217. A likely reason is that the categorical variable in this dataset has only six values and the purchase-rate differences across categories are small. In this setting, basic encoding already captures much of the available category information. This result does not mean target encoding is generally ineffective. Its value is expected to be larger for high-cardinality variables such as brands, stores, channels, or user-source categories [6,8].

Multi-seed ensembling (configuration C) increases F1 only from 0.680 to 0.681. This is consistent with the design goal of this module, which is to reduce random training variation rather than to achieve a large mean performance improvement [10]. Its effect is evaluated more directly through the stability analysis in Section 4.5.

Probability calibration (configuration D) creates an apparently unusual result: F1 falls from 0.681 to 0.626, while the Brier Score improves strongly from 0.217 to 0.192. These two movements are not contradictory. Isotonic regression changes the scale of the probabilities, so the old threshold of 0.5 is no longer well matched to the new distribution. At the same time, the probability values become closer to

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

observed outcomes, which explains the better Brier Score [2,11].

Threshold re-optimization (configuration E) searches again on the calibrated probabilities. F1 returns to 0.681, while the improved Brier Score remains 0.192. This comparison shows why calibration and threshold selection should be completed as consecutive steps. Calibration changes the meaning of the score; threshold selection then adapts the final decision rule to that new score scale.

4.4 Probability Calibration

The Brier Score decreases from 0.2165 before calibration to 0.1922 after calibration, a relative improvement of about 11.2%. This is the clearest gain in the framework that does not depend on choosing a favorable classification threshold. (Table 6)

Table 6. Brier Score Before and After Probability Calibration

Stage	Brier Score	Relative Change
Before calibration	0.2165	—
After calibration	0.1922	−11.2%

The practical meaning of calibration is that model outputs become more useful as probability estimates. For example, if a group of customers receives a calibrated probability around 0.70, the observed purchase rate for similar cases should be closer to 70% than it was before calibration. This property matters for expected-value ranking, marketing budget allocation, and contact decisions, because these tasks depend on the numerical probability rather than only on the ordering of customers [11,12].

A binary decision alone cannot support the same level of differentiation. For example, customers with predicted probabilities of 0.85 and 0.55 may both be classified as purchasers under one threshold, but their expected values can still be very different. Reliable probabilities allow the business to allocate resources with more detail.

4.5 Multi-Seed Stability Analysis

The five single-seed models produce validation F1 scores of 0.6782, 0.6756, 0.6798, 0.6785, and 0.6779. The range is about 0.0042. After averaging the probabilities, the validation F1 is 0.6808, which is higher than all five individual models. (Table 7 and Figure 5)

Table 7. Performance Across Random Seeds and Probability Averaging

Random Seed	Validation F1	Best Iteration	Note
42	0.6782	277	
43	0.6756	87	Early stopping much earlier
44	0.6798	369	
45	0.6785	265	
46	0.6779	261	
Probability average	0.6808	—	Higher than all single models

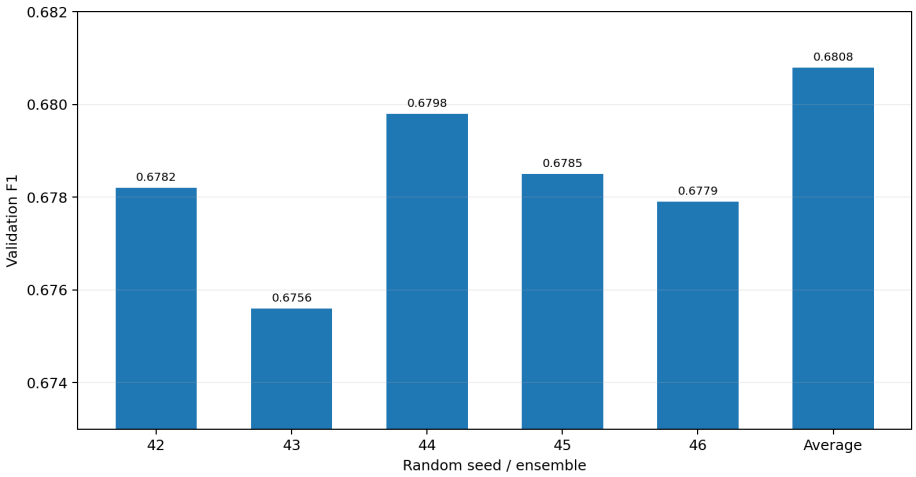


Figure 5. Variation across random seeds and the ensemble result

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

The difference in early stopping rounds is even larger than the difference in F1. The best iteration counts are 277, 87, 369, 265, and 261. Seed 43 stops much earlier than the others. This result indicates that a single reported run can be sensitive to random training conditions [10]. Averaging predictions across multiple seeds reduces this dependence, but training time increases from about 0.58 seconds to 2.98 seconds, close to a five-times increase. Therefore, the method improves stability by accepting additional computational cost.

5. Discussion

5.1 Different Roles of the Framework Components

A main finding of this study is that the modules address different problems, so evaluating them solely with F1 would obscure their true value. Feature expansion and target encoding operate on the input representation. In this dataset, their extra discrimination is limited. Multi-seed averaging operates on training variance and is mainly intended to improve stability. Probability calibration operates on the scale of the output score and improves the meaning of probabilities without trying to improve ranking. Threshold optimization operates on the final decision rule and selects a working point for a given objective.

This separation helps prevent two common interpretation problems. The first is describing a recall increase caused by a lower threshold as a large improvement in model learning. The second is concluding that calibration fails because F1 decreases when the old threshold is kept.

5.2 How the Framework Should Be Valued

The results do not support describing Enhanced LightGBM as a model that substantially improves classification performance. Enhanced and Improved LightGBM have the same F1 score, and the AUC improvement over the best baseline is less than 0.004. The stronger contribution is found in three areas. First, the Brier Score decreases by 11.2%, which makes probability outputs more reliable [2]. Second, multi-seed averaging reduces random variation and improves reproducibility [10]. Third, the leakage-control procedure makes the offline evaluation closer to the information conditions of real deployment [7].

In practical e-commerce applications, prediction outputs are often used for ranking, customer segmentation, and budget allocation rather than treated as final yes/no decisions. In these cases, reliable probabilities can be more useful than a very small increase in classification accuracy. For this reason, the framework is better understood as improving output reliability and decision support.

5.3 Limitations and Future Work

This study has several limitations. First, the main train-validation-test split is created once with random seed 42. The reported test metrics therefore come from one fixed split, and repeated experiments are not used to report means, standard deviations, or confidence intervals [10]. Future work should use repeated stratified cross-validation or repeated train-test splits and apply paired statistical tests when comparing models.

Second, the categorical variable has low cardinality, so the potential advantage of target encoding is not fully tested [6,8]. Future studies should examine data with brands, stores, channels, or user sources that contain many categories. Third, the isotonic calibrator and the classification threshold are both selected using the same validation set. This can introduce some selection bias. A separate calibration set or cross-calibration procedure would provide a more conservative evaluation [2]. Finally, the dataset does not include complete timestamps for all behavior variables. Without full timing information, it is difficult to prove that every behavior happened before the purchase decision. Future work should use time-based data splits whenever timestamps are available [7].

6. Conclusion

This paper develops an Enhanced LightGBM framework for e-commerce purchase prediction by integrating domain-informed feature expansion, leakage-free target encoding, multi-seed probability averaging, isotonic probability calibration, and threshold optimization. The results show that the main

ISSN 2616-5775 Vol. 9, Issue 8: 79-89, DOI: 10.25236/AJCIS.2026.090810

value of the framework is not a large improvement in discrimination, but better probability reliability, training stability, and more appropriate decision-making under different precision-recall trade-offs. Probability calibration improves the interpretability of predicted purchase probabilities, while multi-seed averaging reduces the influence of a single random training run. Overall, the study demonstrates how several existing techniques can be organized into a leakage-aware and reproducible pipeline that better supports practical marketing decisions based on reliable probability estimates.

References

[1] Esmeli, R., Bader-El-Den, M., & Abdullahi, H. (2022). An analyses of the effect of using contextual and loyalty features on early purchase prediction of shoppers in e-commerce domain. Journal of Business Research, 147, 420–434. https://doi.org/10.1016/j.jbusres.2022.04.012

[2] Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. Machine Learning, 112(9), 3211–3260. https://doi.org/10.1007/s10994-023-06336-7

[3] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35, 507–520. https://doi.org/10.52202/068431-0037

[4] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

[5] 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), 78. https://doi.org/10.32604/cmc.2025.076329

[6] Pargent, F., Pfisterer, F., Thomas, J., & Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Computational Statistics, 37(5), 2671–2692. https://doi.org/10.1007/s00180-022-01207-6

[7] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

[8] Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. ACM SIGKDD Explorations Newsletter, 3(1), 27–32. https://doi.org/10.1145/507533.507538

[9] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31, 6638–6648.

[10] Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., Ebrahimi Kahou, S., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., & Vincent, P. (2021). Accounting for variance in machine learning benchmarks. Proceedings of Machine Learning and Systems, 3, 747–769.

[11] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 694–699. https://doi.org/10.1145/775047.775151

[12] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. Proceedings of the 22nd International Conference on Machine Learning, 625–632. https://doi.org/10.1145/1102351.1102430

[13] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

[14] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[15] Alibaba Cloud Tianchi. (2025). Social e-commerce dataset [Data set]. Alibaba Cloud. https://tianchi.aliyun.com/dataset/215680

[16] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432