Journal of Electronics and Information Science (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2025.100209
ISSN 2371-9524 Vol. 10 Num. 2

MDBIF: A Multi-Dimensional Feature and Boosting Integration Framework for O2O Coupon Redemption Prediction

Jiaqi Xu[1,a,#], Zichen Liang[2,b,#]

[1]Lanzhou Zhicheng Academy, Lanzhou, Gansu, China
[2]The High School Attached to Northwest Normal University, Lanzhou, Gansu, China
[a]1059978462@qq.com, [b]1259435499@qq.com
[#]Co-first author

Keywords: O2O Coupons, Machine Learning, Feature Engineering, Model Integration, Targeted Delivery

Abstract: The low redemption rate of coupons in Online-to-Offline (O2O) platforms poses a key challenge for marketing efficiency. To address this, we propose a Multi-Dimensional Feature and Boosting Integration Framework (MDBIF) that captures user, merchant, coupon, and interaction behaviors across seven feature groups with 26 new features. Using mutual information for feature selection and comparing XGBoost, LightGBM, and CatBoost, our framework enhances prediction robustness via data fusion and tuning. Experiments on a real-world Alibaba Tmall dataset (1.75M records) show that LightGBM achieves the best performance (AUC 0.9961, accuracy 0.9815). Key features such as user-specific coupon receipt frequency and merchant distance prove critical. Based on this, we offer actionable targeting strategies to improve O2O coupon effectiveness. Our approach provides a scalable solution for precision marketing in O2O ecosystems.

1. Introduction

In today's digital era, the rapid development of the Online-to-Offline (O2O) business model has made coupons a widely adopted and effective marketing tool, especially on e-commerce platforms such as Alibaba Tmall. These coupons are designed to stimulate consumer behavior and increase merchant sales. However, a critical issue in O2O coupon deployment is the persistently low redemption rate. According to real-world data, a large proportion of distributed coupons remain unredeemed, resulting in wasted resources and reduced marketing efficiency. This problem stems not only from the complexity of user behavior—such as the influence of geographical distance, consumption habits, and coupon types—but also from the challenge of accurately predicting whether a user will redeem a coupon within a limited period (e.g., 15 days), which is crucial for targeted delivery.

To address this challenge, existing research has mainly focused on two directions: theoretical analysis and algorithmic modeling. Theoretically, studies have explored coupon characteristics (e.g.,

discount types and thresholds) and implicit mechanisms of redemption behavior [1-4]. On the algorithmic side, models such as logistic regression, random forests, GBDT, and XGBoost have been widely used to predict coupon redemption outcomes [4-7]. Although these approaches have shown promising results, several limitations remain. First, many studies rely on simplistic feature engineering, often focusing on basic statistics or user-merchant distances, while overlooking multi-dimensional interactions—such as the triadic relationships among users, coupons, and merchants—thus limiting the model's ability to capture complex behavioral patterns. Second, model selection is often restricted to a single algorithm or naive ensemble methods, lacking a systematic comparison and optimization of multiple boosting models. These issues hinder predictive accuracy, especially in scenarios involving imbalanced and time-sensitive data, which are common in O2O environments.

To overcome these limitations, we propose a comprehensive approach that combines multi-dimensional feature engineering with integrated optimization across three boosting models: XGBoost, LightGBM, and CatBoost. Our framework aims to enhance both feature interaction expressiveness and model adaptability. However, achieving this requires addressing two key challenges: (1) how to systematically extract rich, high-level interaction features in O2O scenarios while mitigating noise and redundancy in high-dimensional data; and (2) how to optimize model integration under class imbalance and temporal constraints, ensuring effective fusion without excessive computational overhead.

To address the first challenge, we design seven feature groups covering users, merchants, coupons, and their interactions, generating 26 new features through behavioral statistics (e.g., user-specific coupon redemption rates). Mutual information-based feature selection is employed to reduce dimensionality while preserving relevance. For the second challenge, we employ data fusion and parameter tuning techniques to systematically compare the performance of the three boosting models, ultimately integrating them with LightGBM to enhance robustness on imbalanced data.

In this study, we develop the Multi-Dimensional Feature and Boosting Integration Framework (MDBIF), which constructs interaction-aware representations and integrates advanced boosting models to overcome the limitations of prior work. Experiments on the Alibaba O2O coupon dataset (1.75 million records) show that our approach significantly improves performance, with LightGBM achieving the highest AUC (0.9961), outperforming both XGBoost (0.9958) and CatBoost (0.9957).

Our key contributions are summarized as follows:

1) We propose a comprehensive feature engineering framework to capture complex user-merchant-coupon interactions and address the limitations of shallow feature design.

2) We conduct an in-depth, task-specific comparison and integration of three popular boosting models, filling a gap in multi-model ensemble research for O2O prediction tasks.

3) Our framework demonstrates strong empirical performance and provides practical targeting strategies, including distance- and volume-based coupon delivery recommendations.

The remainder of this paper is structured as follows: Section 2 reviews related work; Section 3 introduces our proposed method; Section 4 presents experimental results and analysis; Section 5 concludes the paper.

2. Related Work

Feature Engineering. With the growing popularity of the Online-to-Offline (O2O) business model, many studies have been devoted to understanding and predicting coupon usage behaviors. In the context of O2O coupon redemption prediction, feature engineering plays a critical role in determining model performance. The goal of feature engineering is to transform raw data into high-informative features that capture user behavior, merchant characteristics, and coupon strategies. Huangbo et al. [1] demonstrated that features such as user-merchant distance, merchant-level sales

volume, and the total number of redemptions for various coupon types have a significant impact on prediction accuracy. Other studies [2,3,4] have proposed a systematic approach to feature construction across multiple dimensions, including user profiles, merchant attributes, coupon details, user-merchant interactions, and contextual factors, providing a transferable paradigm for feature engineering in this task. To capture behavioral dynamics, Wang, L. (2023) [7] introduced a sliding window mechanism to generate temporally-aware training samples, addressing the lack of temporal structure in raw features. Song, X. (2021) [4] employed a hybrid method combining mutual information and recursive feature elimination using Random Forests was employed to select high-contribution features, thereby reducing feature dimensionality and training complexity. Furthermore, Song, X. (2021) [4] addressed class imbalance in coupon redemption prediction using a BG-XGBoost algorithm, achieving notable improvements over baseline models.

Prediction Models. Machine learning models remain the predominant approach for predicting O2O coupon redemption. Prior studies [5,6,2,3] commonly frame this task as a binary classification problem. Several works have conducted systematic comparisons of classic models, such as logistic regression, random forests, and XGBoost [2,3], while others have explored progressive modeling strategies to refine prediction quality [8]. To further enhance performance, some studies investigated ensemble techniques, including Stacking and Balanced Bagging, leveraging complementary strengths across models to improve generalization [7,4].

3. Methodology

3.1 Data Exploration

We conducted exploratory data analysis (EDA) using Python on the original dataset, which contains a total of 1,754,884 records. Each coupon in the dataset has a validity period of 15 days. Preliminary statistics reveal that 1,053,282 records involve users who received coupons, but only 96,524 of them redeemed the coupons within the valid period. This indicates that the vast majority of distributed coupons were not used, highlighting the low coupon redemption rate. Such findings underscore the practical importance of accurately identifying potential users and optimizing coupon targeting strategies for merchants.

3.2 Data Preprocessing

Based on our analysis of the key factors affecting prediction performance, we focused on preprocessing two critical fields: “Discount_rate` and `Distance”.

For the “Discount_rate” field, we first extracted the threshold amount required for a discount (denoted as “discount_man”) and the corresponding reduction amount (“discount_jian”) for conditional coupons. A custom function “getDiscountType” was then used to categorize the discount types into three classes: no discount (−1), conditional discount (1), and direct discount (0). This transformation produced a new feature, “discount_type”, which enhances the model's ability to distinguish between different discount mechanisms.

Regarding the “Distance” field, we found that missing values were recorded as the string “null”. To retain the semantic meaning of "unknown distance" while ensuring data type consistency, these `'null'` entries were replaced with −1, and the entire field was converted to integer type. These preprocessing steps significantly improved the structure and consistency of the data, laying a solid foundation for subsequent feature engineering and modeling.

3.3 Feature Engineering

To enhance the model's capacity to capture complex behavioral patterns in O2O coupon redemption, we construct seven feature groups covering users, merchants, coupons, and their interactions. These features are derived through statistical aggregation and behavioral analysis. A total of 26 new features are designed as follows: (Table 1-7)

Table 1 Merchant feature group.

Feature Name	Description
Merchant_id	Unique identifier for each merchant.
merchant_coupon_total	Total number of coupons issued by the merchant.
merchant_coupon_use_times	Number of times the merchant's coupons were redeemed.
merchant_min/max/mean_distance	Minimum, maximum, and average user distance from the merchant.
merchant_coupon_rate	Redemption rate of the merchant's coupons.

Table 2 User Feature Group.

Feature Name	Description
User_id	Unique identifier for each user.
user_coupon_get/use_times	Number of coupons received / redeemed by the user.
user_min/max/mean_distance	Minimum, maximum, and average distance between user and merchant.
user_coupon_rate	Overall coupon redemption rate for the user.

Table 3 Coupon Feature Group.

Feature Name	Description
Coupon_id	Unique identifier for each coupon.
coupon_total	Total number of such coupons issued.
coupon_use_times	Number of times this coupon type was used.
coupon_discount_rate	Discount value of the coupon.
coupon_rate	Redemption rate of this coupon type.

Table 4 User–Coupon Interaction Group.

Feature Name	Description
user_specific_coupon_rate	Redemption rate of a specific coupon by a specific user.
user_specific_coupon_use_times	Number of times a specific user used a specific coupon.

Table 5 User–Merchant–Coupon Interaction Group.

Feature Name	Description
merchant_specific_coupon_total	Number of times a specific coupon was issued by a specific merchant.
merchant_specific_coupon_use_times	Number of times the merchant's specific coupon was redeemed.

Table 6 Merchant–Coupon Interaction Group.

Feature Name	Description
user_merchant_coupon_rate	Probability of a user redeeming a coupon at a specific merchant.
user_merchant_coupon_use/get_times	Number of times a user redeemed/received a specific coupon from a merchant.
merchant_coupon_for_specific_user_use_rate	Redemption rate of a merchant's coupon for a specific user.

Table 7 Other Feature Group.

Feature Name	Description
Discount_rate	Discount rate of the coupon.
Distance	Distance between user and merchant.
Date_received	Date on which the user received the coupon.
discount_type	Type of discount: none (−1), direct (0), threshold-based (1).
discount_man	Threshold amount required for the coupon to be valid.

These features serve as the input for downstream boosting models, enabling fine-grained modeling of redemption behaviors from multiple perspectives.

4. Experiments

4.1 Evaluation Metrics

To evaluate the performance of our models on the binary classification task of predicting coupon redemption, we consider several standard metrics: Accuracy, Precision, Recall, F1 Score, ROC curve, and AUC (Area Under the Curve). These metrics are computed based on the four basic classification outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Here, "positive" and "negative" refer to the predicted class, while "true" and "false" denote prediction correctness.

We focus primarily on the AUC as the core evaluation metric due to its robustness in imbalanced classification settings. The definitions of the key metrics are as follows:

(1) Accuracy: The ratio of correctly predicted samples to the total number of samples:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (1)$$

(2) Precision: The ratio of correctly predicted positive samples to all samples predicted as positive:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

(3) Recall: The ratio of correctly predicted positive samples to all actual positive samples:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

(4) ROC Curve: A curve that plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. TPR indicates the proportion of actual positives correctly identified, while FPR measures the proportion of actual negatives incorrectly classified as positive.

(5) AUC: The area under the ROC curve, ranging from 0 to 1. A higher AUC indicates a better-performing model.

4.2 Model Construction and Performance Comparison

The core task of this study is to predict whether a user will redeem a coupon based on historical O2O data, formulated as a binary classification problem. We employ machine learning algorithms well-suited for this task, particularly tree-based boosting models. While traditional studies often use XGBoost, GBDT, or Random Forests, we focus on three gradient boosting models: XGBoost, LightGBM, and CatBoost.

We first apply a multidimensional feature engineering process to generate new features, and then train models on two training subsets using the train_test_split function. The training results are fused, and model parameters are tuned using AUC as the primary metric. The experimental results are summarized below: (Table 8)

Table 8 The experimental results.

Model	Accuracy	Precision	Recall	F1 Score	AUC
XGBoost	0.9805	0.8205	0.8710	0.8450	0.9958
LightGBM	0.9815	0.8311	0.8743	0.8521	0.9961
CatBoost	0.9798	0.8213	0.8562	0.8384	0.9957

The results show that all three models achieve excellent performance. Among them, LightGBM achieves the highest AUC (0.9961) and also outperforms the other models in Accuracy, Precision, Recall, and F1 Score. While XGBoost shows competitive results, it slightly lags behind LightGBM. CatBoost, although comparable in terms of score, requires significantly more training time—several times longer than XGBoost and LightGBM—making it less favorable in time-sensitive scenarios.

Overall, LightGBM demonstrates superior performance and computational efficiency in our experiments, making it the most suitable choice for this O2O coupon redemption prediction task.

4.3 Feature Importance Analysis

Feature importance visualization refers to presenting the key features identified by the trained model in graphical or tabular form, allowing better interpretation of the factors influencing coupon redemption. In this study, we analyze feature importance scores extracted from three boosting models—XGBoost, LightGBM, and CatBoost—using their built-in `feature_importances` attributes. We summarize the top 10 most important features for each model and perform a comparative analysis.

Table 9 XGBoost Feature Importance.

Feature	Score	Feature	Score
user_specific_coupon_get_times	0.2770	user_merchant_coupon_get_times	0.0741
coupon_total	0.1436	month	0.0653
merchant_mean_distance	0.1165	merchant_coupon_total	0.0432
discount_rate	0.0969	user_min_distance	0.0361
discount_man	0.0809	user_coupon_get_times	0.0179

Table 10 LightGBM Feature Importance.

Feature	Score	Feature	Score
coupon_total	624	month	202
merchant_mean_distance	509	user_specific_coupon_get_times	184
merchant_coupon_total	459	discount_man	178
discount_rate	252	user_coupon_get_times	152
user_merchant_coupon_get_times	235	user_mean_distance	91

Table 11 CatBoost Feature Importance.

Feature	Score	Feature	Score
merchant_mean_distance	20.8403	user_specific_coupon_get_times	7.3322
merchant_coupon_total	11.2319	month	7.1403
merchant_specific_coupon_total	10.8959	user_merchant_coupon_get_times	6.7711
coupon_total	7.6638	user_min_distance	6.1581
discount_man	7.3844	coupon_discount_rate	4.3795

From the above results (Table 9-11), it is evident that factors influencing coupon redemption behavior span across merchant, user, and coupon dimensions.

- Merchant-related features include: total number of coupons issued, average distance between users and merchants, total sales, 15-day redemption rate of merchant-specific coupons, and coupon usage frequency.
- User-related features include: the number of coupons received per month, user-specific coupon receipt frequency, redemption activity at specific merchants, and 15-day user-merchant-coupon interaction rate.
- Coupon-related features include: the total number of issued coupons per type and the usage count for each type.

4.4 Recommendations for Targeted O2O Coupon Delivery

This study aims to address two key questions: how to identify target users, and what types of coupons to issue to them. Based on the above feature analysis, we propose the following strategies:

(1) Identifying Target Users. Key indicators such as a user's history of receiving coupons and spending behavior at specific merchants are strong predictors of redemption. Merchants can analyze user consumption patterns—such as frequent visits to certain stores or preferences for specific coupon types—to identify high-potential customers.

We categorize users into three segments:

- High-value users: Frequently receive and use coupons; should be prioritized for regular coupon distribution.
- Mid-value users: Receive coupons but rarely redeem; require tailored strategies such as adjusting frequency or type of delivery to improve engagement.
- Potential users: Do not receive coupons but make purchases; could be activated through broader targeting or high-value incentives.

For mid- and potential users, strategies may include increasing perceived coupon value, offering diverse coupon types, or designing personalized delivery based on behavioral clustering to elevate their engagement levels.

(2) Designing Effective Coupons The type of coupon significantly affects user redemption behavior. Preferences differ between discount-based and threshold-based coupons. Merchants can leverage purchase history and product characteristics to determine the most suitable discount format, thereby increasing the perceived value of the coupon.

In addition, setting an appropriate validity period is essential. By analyzing when coupons are most likely to be redeemed, merchants can optimize the time window to maximize usage while minimizing waste and operational costs.

(3) Merchant-Centric Delivery Strategies The user–merchant distance and total sales volume are key determinants of coupon redemption. Merchants should adopt location-sensitive delivery strategies:

- Offer lower-value coupons to nearby users to maintain engagement cost-effectively.
- Offer higher-value coupons to distant users to incentivize longer travel and increase store reach.

Such distance-adjusted strategies can improve both redemption rates and store visibility, enhancing long-term customer acquisition.

5. Conclusion

This study proposes a Multi-Dimensional Feature and Boosting Integration Framework (MDBIF) to improve O2O coupon redemption prediction. By constructing 59 features across seven user–merchant–coupon interaction groups and applying mutual information selection, the framework enhances behavioral modeling. Experiments on the Alibaba Tmall dataset show that LightGBM outperforms XGBoost and CatBoost, achieving the highest AUC (0.9961) and accuracy (0.9815). Key predictive features include user-specific coupon activity, merchant distance, and coupon volume. Based on these findings, we offer targeted delivery strategies to improve redemption efficiency. Future work may explore deep learning and real-time data adaptation for dynamic O2O environments.

References

[1] Huang, B., Hu, X., & Zhai, J. (2023). Research on O2O coupon distribution strategy based on machine learning. In Proceedings of the 18th Annual Conference of the Chinese Academy of Management Science and the 10th Anniversary

Forum of the Belt and Road Initiative (pp. 511–521). China University of Mining and Technology; Hunan University. https://doi.org/10.26914/c.cnkihy.2023.029347
[2] Zhang, X., Qiu, J., & Li, B. (2024). Precise Issuance of Meituan Merchants' Coupons with Machine Learning. In Proceedings of the International Conference on Machine Learning, Pattern Recognition and Automation Engineering (pp. 71-75).
[3] Mei, H., & Li, X. (2019). Research on e-commerce coupon user behavior prediction technology based on decision tree algorithm. International Core Journal of Engineering, 5(9), 48-58.
[4] Song, X. (2021). Research and application of O2O coupon usage prediction based on XGBoost (Master's thesis). Southwest Jiaotong University. https://doi.org/10.27414/d.cnki.gxnju.2021.002056
[5] Weihan Yang,Zijie Zhang,Runze Liu & Jinjiang Liu.(2023).Forecast of O2O Coupon Consumption Based on XGBoost Model.Academic Journal of Computing & Information Science,6(8)
[6] Qiongyu, S. (2020). Prediction of o2o coupon usage based on xgboost model. In Proceedings of the 2020 11th International Conference on E-business, Management and Economics (pp. 33-36).
[7] Wang, L. (2023). Research on O2O coupon prediction based on data mining (Master's thesis). Dalian University of Technology. https://doi.org/10.26991/d.cnki.gdllu.2023.004622
[8] Meijuan, S., & Kaili, Y. (2024, December). Research on O2O Coupon Usage Prediction Based on Machine Learning. In 2024 IEEE 2nd International Conference on Electrical, Automation and Computer Engineering (ICEACE) (pp. 1637-1641). IEEE.