A Multi-Objective Genetic Algorithm-Optimized LightGBM Framework for Customer Segmentation and Strategy Optimization in Cross-Border E-Commerce

Zhao Xu[1*], Xing Yan[2]
[1]School of Business and Management, Wuxi Normal College, Wuxi, Jiangsu, 214153, China
[2]Taizhou Institution of Sci. & Tech., Njust, Taizhou, Jiangsu, 225300, China
E-mail：wxncxuzhao@126.com, audit883@126.com

Keywords: cross-border, e-commerce, strategy optimization, moga, lightgbm, precision marketing, customer segmentation.

Received: July 15, 2025

Cross-border e-commerce has grown rapidly in the context of changing global trade, necessitating the urgent need for more clever and flexible consumer segmentation techniques to improve strategic operations and precision marketing. When it comes to managing high-dimensional, noisy, and behaviourally varied client data, traditional segmentation strategies often fail. In order to bridge this gap, this work suggests a unique hybrid technique that optimizes client segmentation and personalised strategy suggestion by combining the Light Gradient Boosting Machine (LightGBM) with a Selective Multi-Objective Genetic Algorithm (MOGA). To address concerns of incompleteness, duplication, and inconsistency, the technique entails gathering customer contact, order, and product data from a cross-border cosmetics e-commerce platform. Based on the data of 7000 customer purchases on a cosmetics online site. They are frequency, monetary value, product types and time-purchase trends. The data were divided into 70 percent training data, 15 percent validation and 15 percent testing data after pre processing. The proposed MOGA-LightGBM model aims at optimizing the accuracy, F1-score, and ROI by simultaneously tuning the hyper parameters and feature selection. This is followed by a thorough preparation of the data. The suggested MOGA-LightGBM model optimizes a number of factors, including accuracy, recall, and campaign efficiency, in order to maximize classification performance and marketing ROI at the same time. According to experimental data, the model performs better than benchmark methods (RFM + K-Means, XGBoost) with an accuracy of 93%, an F1-score of 85%, and a strategy ROI improvement of 17.4%. This study offers a scalable, data-driven framework for precise operations in cross-border e-commerce and highlights the possibilities of evolutionary optimization in consumer analytics.

Povzetek: Hibridni okvir MOGA-LightGBM je razvit za segmentacijo kupcev in optimizacijo trženjskih strategij v čezmejni e-trgovini. Večciljna genetska optimizacija izboljša točnost, F1-mero in donosnost (ROI) na realnih podatkih.

1 Introduction

The internet's quick development and rising disposable incomes have made online buying a necessity in modern living, drastically changing the dynamics of conventional commerce. Community e-commerce platforms (CECPs) have significantly increased product accessibility and expedited the procurement process, and the COVID-19 epidemic has helped to accelerate online-to-offline (O2O) commerce [1]. The number of transactions in China's O2O market is predicted to reach 250 billion yuan in 2023, demonstrating strong development and underscoring the significance of comprehending the variables affecting customer loyalty and repurchase intentions in this area. Retaining consumers and increasing their repurchase intentions remain major issues [2], even with the widespread deployment of CECPs, which are distinguished by their immediacy and ease. Less attention has been paid to how particular characteristics of CECPs affect perceived value and repurchase intentions, especially for everyday requirements. Previous research has concentrated on the effects of overall website and product quality on customer behaviour [3]. These items provide a unique perspective for analysing customer behaviour on CECPs since they are necessary and often bought.

The term "cross-border e-commerce" particularly refers to businesses that use cross-border electronic commerce platforms, which include both self-built and third-party platforms. Cross-border e-commerce is the network hub of transaction activities in cross-border e-commerce transactions. It serves as a bridge between the supply and consumption of commodities and is not only a medium for commodity browsing and display but also a place for commodity trading [4]. When trading commodities, cross-border e-commerce will be very difficult and complex, leading to a variety of alternative ways for it to operate. The relevant departments of cross-border e-commerce should concentrate on the issue of how to run cross-border e-commerce in a reasonable and efficient manner to better serve consumers in the future [5].

Cross-border e-commerce is becoming an increasingly important component of China's international trade as a result of the wave of informatisation, which has altered both the old modes of consumption and trade. The ability to effectively classify consumers in order to deliver personalised services has become crucial for e-commerce businesses due to the industry's fast expansion [6]. Every business is vying for greater consumer resources in the highly competitive e-commerce sector. Customers' wants and behaviours may be better understood by businesses via customer segmentation, which enables them to implement customised marketing techniques that increase sales and customer happiness. Conventional manual categorisation techniques are unreliable, ineffective, and unable to handle the demands of extensive data processing. We can increase sales efficiency and customer satisfaction by better understanding the traits and requirements of client groups via the analysis and mining of customer data [7,8].

However, the competition between e-commerce companies is fiercer now than it was in the past few decades due to the large number of businesses that have crowded into this market. This is also a major factor in traditional commerce realising that digital commerce is a trendy way to increase profits and transferring their commerce model [9]. Because they have a lot of options, suggestions will be crucial when customers are making a selection. Therefore, in order to increase customer loyalty and encourage them to make more purchases, it is essential to provide them with correct suggestions. This entails suggesting items that they are most likely to purchase. Conventionally, recommendations are made based on clients' past purchases, which is not only illogical but also inefficient [10]. The recommendation system will be enhanced in this article via the usage of market segmentation.

Research questions lead to this study:

RQ1: Does a hybrid MOGA-LightGBM model provide better and practical customer segmentation than the standard practices (e.g., RFM-KMeans, or standalone boosting)?

RQ2: Can multi-objective genetic algorithm optimization of LightGBM parameters enhance marketing performance (ROI) over and above the base strategies?

Hypothesis (H1): A MOGA-LightGBM model will be able to enhance the accuracy of the segmentation and F1-score and ROI will increase by at least X percent in comparison with the highest-performing domestic model.

Contribution of this study:

This research significantly advances the area of cross-border e-commerce by presenting a new hybrid model that combines a Light Gradient Boosting Machine (LightGBM) and a Selective Multi-Objective Genetic Algorithm (MOGA) for accurate consumer segmentation and strategy suggestion. The suggested MOGA-LightGBM architecture concurrently maximises many goals, such as classification accuracy, recall, and business return on investment (ROI), in contrast to conventional clustering or single-objective optimisation techniques. When the model is used to handle complicated, high-dimensional client data from a cosmetics e-commerce platform, it shows great practical relevance and scalability. The experimental findings demonstrate that the suggested method works noticeably better than current techniques, with a 93% accuracy rate and a discernible boost in strategy ROI. Additionally, the model makes it possible to precisely identify consumer subgroups like deal-seekers, high-value clients, and inactive users, offering useful information for tailored marketing. Finally, the research offers a wide range of user-friendly visualisations, including radar plots and grouped bar charts, that help both technical and commercial stakeholders understand the model's findings. All things considered, the study offers a solid, understandable, and practical approach that enhances precision marketing's potential in the global e-commerce environment.

2 Literature review

A key internationalisation strategy for businesses is the way of entry into a foreign market, which has a significant influence on the commercial success of the company, according to literature [11]. E-commerce across borders is a very sensible option. According to literature [12], Chinese e-commerce companies have discovered a great chance to market their goods abroad under the Belt and Road program. While there are undoubtedly numerous distinctions between domestic and international e-commerce, there are also similarities in terms of release and marketing tactics. The operating style of cross-border e-commerce has been particularly popular in the sphere of international trade, according to literature [13]. This is because of the network's quick growth, globalisation, ease, and mobility. It was discovered that creating individualised and focused marketing campaigns for cross-border e-commerce may raise businesses' overall marketing proficiency. According to studies in the literature [14], cross-border e-commerce platforms must provide services and marketing plans that are tailored to the needs of various user groups in order to increase customer satisfaction and the rate at which purchases are made. Additionally, e-commerce platforms have both opportunities and challenges as a result of the shifts in cross-border e-commerce, including platform fission and marketing innovation.

Table 1: Summary on related works

Ref	Objective	Dataset	Method Used	Performance Metrics	Limitations / Gaps	How MOGA-LightGBM Fills Gap
[15]	Churn prediction with feature reduction	Telecom & generic churn datasets	Multiple ML models + feature selection	Accuracy, recall, ROC-AUC	Focus only on churn, no segmentation; no multi-objective optimization	MOGA-LightGBM integrates feature selection + segmentation + hyperparameter tuning simultaneously
[16]	Predict customer loyalty	Finance industry dataset	LightGBM	Accuracy, F1-score	Single objective; not multi-objective or GA-based; domain-specific	Adds multi-objective GA layer to LightGBM and adapts to multiple domains
[17]	Big data classification & mining in AI era	Large e-commerce data	AI-enhanced classification	Precision/recall	High-level AI discussion, lacks concrete hybrid models for segmentation	MOGA-LightGBM offers concrete, trainable hybrid for segmentation and recommendation
[18]	E-commerce user segmentation & marketing strategy	E-commerce platform data	Cluster analysis	Segmentation quality	Traditional clustering, no gradient boosting integration	Combines segmentation with predictive model inside one pipeline
[19]	Time-series clustering of customer behavior	Customer time-series	Genetic algorithm + data mining	Clustering quality	Focused on clustering only; no downstream prediction	MOGA-LightGBM unifies GA-based segmentation with predictive LightGBM
[20]	Model combination for cross-border e-commerce	Cross-border platform	Robot hybrid algorithm	Platform performance	Lacks explicit segmentation or multi-objective trade-offs	MOGA-LightGBM explicitly optimizes multiple metrics for segmentation + recommendation
[21]	Analyze comment data on e-commerce platforms (RPA robots)	Customer reviews	RPA + text analysis	Sentiment accuracy	Text only, no integrated predictive model	MOGA-LightGBM can incorporate textual features alongside transactional ones

Key justification:

Each of the baselines has segmentation, feature selection, and prediction as an independent decision or is limited in its optimization. No one combines a multi-objective GA with LightGBM to select features/segments simultaneously as well as tune hyperparameters and achieve multiple performance objectives (e.g., accuracy, ROI, F1). MOGA-LightGBM transparently addresses this gap by (1) dealing with heterogeneous features (RFM, time-based, categorical, textual) (2) jointly optimizing many objectives and (3) offering a single, scalable structure of personalized recommendations and marketing.

3 Methodology

3.1 Data acquisition

Due to the rapid growth of e-commerce in recent years, several e-commerce websites have amassed a substantial quantity of data, including information on customers, sales, commodities, etc. These statistics are essential when businesses are creating their marketing or sales plans. Our success also depends on how to get useful client segmentation data. The information utilised in this study was taken from an online store selling cosmetics. We must extract the needed data from the database, such as the customer information table, commodity information table, customer order table, etc., since the data source contains a lot of complicated data.The data is a collection of a seven thousand transaction and customer records of an online cosmetics store (6,782 records after cleaning). It was divided into 70 percentage train/validation/test split to maintain the distribution of the customer segments. In model development, the tuning of hyperparameters was done within the MOGA framework and inner loop tuned the validation set to determine which LightGBM configurations should be used by evaluating their fitness on the validation set, and the outer test set was not used to introduce bias during the estimation of final performance.

3.2 Data preprocessing

The original data often has flaws including incompleteness, repetition, and nonstandardity. To make the data as consistent as feasible, the original data may be restored by data cleaning techniques including missing value processing, isolated point exclusion, noisy data elimination, etc. Certain choices in the corporate database are optional when customers create their accounts, and some of these options may include sensitive information. As a result, consumers may be hesitant to fill out information, leaving the database with many empty entries. As a result, we must address the missing values before we can analyse the data. Typical techniques for handling missing values include approximated filling, manual processing, and so on. Repeated, inaccurate, and incomplete data are referred to as noise in data. Statistical concepts may be used to identify error data. In general, data that exceeds the mean value by two positive and negative standard deviations might be considered noisy data. Data that has missing information is called incomplete data. For instance, some customers' frequently used language information is incomplete and might be considered noise data. To put it simply, duplicate data is information that has been duplicated. When a customer's consumption patterns are captured twice, it will undoubtedly have an incorrect impact on the analysis that follows. The attribute types for each customer record with several dimensions are inconsistent; some are text, some are Boolean, and others are numeric. The source records must be transformed to comply with data mining criteria in order to enable the subsequent accurate computation.

3.3 Developing of customer segmentation model

Establishing a subdivision model, or which subdivision technique may provide superior subdivision outcomes, is another crucial step before implementing a subdivision algorithm. Divide the customer groups according to the subdivision technique after figuring out the attribute index of the subdivision, and then extract the group characteristics for each customer group. After pre-processing the customer data, a segmentation model is created based on the pertinent customer attributes. The data is then combined and separated using the selective MOGA-LightGBM integration algorithm. The segmentation results are then used to divide the various customer groups, and group characteristics are extracted to provide relevant marketing recommendations.

- Data preprocessing objective: Process raw e-commerce data (customer, product, and order tables) to obtain a high-quality dataset by cleaning, normalizing, and transforming raw data and dealing with missing, duplicate, or noisy data to ensure high-quality end analysis.
- Segmentation objective: Segmentation: With the help of the cleaned data, distinguish individual customer segments with discernible behavioral and value-based profiles, which can be used to create individually tailored marketing approaches.
- MOGA design goal: Find Pareto-optimal solutions to the design issues of LightGBM by optimizing both hyper-parameters and features selection simultaneously on multiple design objectives, accuracy, F1-score, and ROI with the help of a multi-objective genetic algorithm.
- Performance Criterion: The model will be declared successful when it has an improvement in ROI of ≥17.4% and segmentation accuracy/

AUC of ≥15% compared to the strongest baseline method.

3.4 Multi-objective genetic algorithm (MOGA)

The evolutionary search for optimal solutions in nature served as the inspiration for genetic algorithms, which are computer algorithms. Because it uses mathematical models with high accuracy values and can handle a variety of issues with complicated search spaces, this approach has been employed extensively. Therefore, it was deemed appropriate to use genetic algorithms in a variety of domains, particularly when producing forecasts for future events, such as stock price projections, currency exchange rates, marketing strategy recommendations, and cross-border e-commerce, or when creating predictions for customer segmentation. Initialising individuals or creating individuals from a random collection of genes (chromosomes) is the first step in genetic algorithms. This chromosome held the solution to the problem. Reproduction, which involves a crossover and mutation process to boost the population, came next. How likely a chromosome was to be a solution is shown by its fitness value; the higher the deal, the more likely it was. To determine this fitness value, evaluation was a necessary step. Members of the spawn and population set are chosen in the last stage, selection.

3.4.1 Population initialization

The crucial phase in genetic algorithm prediction was often identifying the finest historical data patterns technique, regression. The goal of this approach is to identify a pattern that closely reflects the characteristics of previous data on China's inflation rate. In this research, the Gradient Boosting Machine (GBM) was used to find a design that best matched the features of China's historical rates. Boosting functions (1) were used to represent the pattern, just like in the equation. This function would be used during initialisation to generate a prediction model utilising training data.

$$z = \theta_0 + \theta_1 w_1 + \theta_2 w_2 + \theta_3 w_3 + \cdots + \theta_m w_m$$
(1)

Where:

z : The Consumer segmentation Price Index for a month A was predicted

$w_1 \ldots w_m$: Index of Consumer segmentation Prices for Months A1 to An

$\theta_0 \ldots \theta_m$: The use of random numbers to represent each gene on a chromosome

3.4.2 Calculation of fitness value

The actual transaction and Mean Square Error (MSE) are combined to get this study's fitness score (f). When the ideal value is reached with the lowest MSE value, the fitness value will be greater. There would be a prediction inaccuracy using equation (2).

$$f = 1/(MSE + 0)$$
(2)

To anticipate the china's inflation rate, one must first calculate the MSE (value, square all the data errors, and then divide by the overall number of mistakes. The equation was used to compute the MSE. (3).

$$MSE = \frac{1}{m}\sum_{j=1}^{m}(Z_j - Z_j')^2$$
(3)

Where:

m : The number of data
Z : Prediction of data
Z' : Ground truth of data

3.4.3 Crossover

By using the crossover procedure, which involves combining pieces of the alela line's DNA genome to create a hereditary genome or progeny, new individuals were created within a generation. The crossover procedure makes use of the specified alpha values and the whole-arithmetic method. When changing the population size, the risk of crossover (Pc) will be taken into account. The following formula is used to generate a random gene selection for the crossover process:

$$child\ 1 = \alpha . w_j + (1-\alpha). z_j, 1 \le j \le m \tag{4}$$

$$child\ 2 = \alpha . w_j + (1-\alpha). z_j, 1 \le j \le m \tag{5}$$

3.4.4 Mutation

The alteration procedure was performed to individuals after parent cross or crossover findings. This approach changed the ability of one or more genes in a population to delay premature convergence, which is the achievement of a value or performance before it has reached its maximum potential. The number of participants in the mutation process is determined by the calculated probability of the mutation (Pm). MO programming had several goals, none of which could be maximised simultaneously. Consequently, the decision-makers search for the optimal choice. In MO programming, optimality was replaced by efficiency or Pareto optimality. According to the Pareto optimum solution, it was feasible to improve one objective function while lowering at least one of the other objectives. One technique for dealing with multiobjective problems was the ε-constraint, which views one of the goals as the dominating objective function. While the ε-constraint approach handles secondary objectives as constraints, the primary goal was optimised.

$$minE_1(w)$$
$$Subject\ to\ E_2(w)\leq\varepsilon_2 E_3(w)\leq\varepsilon_3 \ldots E_O(w)\leq\varepsilon_O \tag{6}$$

The -constraint strategy required determining the lowest and maximum values for each objective, where O was the number of competing goals. For this, the payout table method was often used. Objective function i (i ¼ 1, …,O) was reduced to a single-goal issue, and values for the objective functions were entered in the relevant columns of the ith row in order to calculate the value, O rows, and O columns of the payment table. Thus, using the lowest and greatest values in this table's ith column, the range of the objective function I was determined.

$$minE_1(w)$$
$$Subject\ to\ E_2(w)\leq\varepsilon_{2,j} \tag{7}$$

$$\varepsilon_{2,j} = max(E_2) - \left[\frac{max(E_2)-min(E_2)}{iter_{max}}\right] iter\ \ iter = 0,1 \ldots, iter_{max} \tag{8}$$

$E1(w)$ and $E2(w)$ were the objective functions for recommendation strategies and feature extracted, respectively. $iter$and $inter_{mix}$were the maximum number of iterations and the current iteration, respectively. $min(E2)$and $max(E2)$, the payment table's lowest and maximum values for extracted segmentation, respectively.

$$\mu_1^{iter} = \begin{cases} 1 & E_j^{iter}<min(E_j) \\ \dfrac{max(E_j)-E_j^{iter}}{max(E_j)-min(E_j)} & <min(E_j)<E_j^{iter}<max(E_j) \\ 0 & <E_j^{iter}>max(E_j) \end{cases} \tag{9}$$

$$\mu^{iter} = min\left(\mu_1^{iter}, \ldots . \mu_o^{iter}\right) \tag{10}$$

A popular heuristics algorithm used to address engineering issues GA. Algorithm 1 was the first algorithm developed in MOGA that combines while preserving population variety and fitness sharing.

Algorithm 1: Multi-objective genetic algorithm (MOGA)

```
Step1: Initialize
        Select a random initial population $O_s \subset T$
Step 2: Begin
    While (stopping criteria are NOT satisfied) do
        Evaluate fitness of the population
            Select parents using a GBM method
            Apply crossover and mutation,
            Update $O_s, MC$
            Set $s = s + 1$   (20)
    End while
        Return $MC$ and $O_s$
End
```

The Multi-Objective Genetic Algorithm (MOGA) is used in this research as an intelligent optimisation framework to improve cross-border e-commerce strategy suggestion and consumer segmentation. Because customer segmentation entails maximising classification accuracy, balancing segment distribution, and enhancing marketing ROI all of which are competing goals are MOGA is especially well-suited for this job. MOGA develops a population of possible segmentation solutions over many generations by simulating the process of natural selection. It uses genetic processes including crossover, mutation, and selection in each iteration to optimise many goals at once and investigate a variety of solutions. When used with the LightGBM model, MOGA aids in choosing the most relevant customer qualities and segment borders that result in excellent classification performance. A collection of Pareto-optimal segmentation models that achieve the best trade-offs between performance indicators are the end result. This maximises both analytical accuracy and economic impact by allowing the system to more precisely identify discrete client groups and connect those segments with successful marketing tactics.

3.5 Light gradient boosting machine (LGBM)

LGBM is a machine learning-based gradient-boosting framework with improved performance. When compared to more conventional boosting algorithms like XGBoost, it is intended to increase efficiency and scalability. By using histogram-based methods, lowering memory use, and utilising a leaf-wise growth approach with depth limitations to avoid overfitting, LGBM improves training time. LightGBM uses four main strategies to maximise decision tree learning. ML speeds up calculations and uses less memory by converting continuous features into discrete bins. Gradient-Based One-Side Sampling (GOSS) increases efficiency without appreciably sacrificing accuracy by randomly sampling minimal-gradient cases while retaining steep-gradient ones. By combining mutually exclusive features, Exclusive Feature Bundling (EFB) lowers memory use and dimensionality. Although a depth restriction is required to avoid overfitting, the leaf-wise growth technique improves model accuracy by choosing the leaf with the maximum information gain for expansion. LightGBM is very effective for big datasets to these optimisations. Training dataset is given by, $W = \{(W_j, z_j)\}_{j=1}^{n}$. The goal is to find an approximate function $\hat{e}(w)$that closely estimates $e^*(w)$to minimize expected values of specific loss operations $(z, e(w))$, objective function mathematical representation as,

$$\hat{e}(w) = \arg \underset{e}{min} F_{z,W} K(z, e(w)) \tag{11}$$

This function minimizes the expected loss operation to optimize predictions. The additive learning process is expressed as,

$$e_S(W) = \sum_{s=1}^{S} e_s(W) \tag{12}$$

where $e_S(W)$denotes the weak learner at iteration s, and $e_S(W)$is the final model after$S$ boosting iterations.

LGBM builds a model as a sum of regression trees.Leaf weight optimization is computed by

$$\omega_i^* = -\frac{\sum_{j\in J_i} h_j}{\sum_{j\in J_i} g_j + \lambda} \tag{13}$$

here, $h_j \; and \; g_j$are the primary and secondary gradients of the loss operation, optimizing the tree's leaf nodes, λ is the regularization parameters .Tree structure optimization (Extreme Values of Γ_S^*) is represented by

$$\Gamma_S^* = -\frac{1}{2}\sum_{i=1}^{I} \frac{(\sum_{j\in J_i} h_j)^2}{\sum_{j\in J_i} g_j + \lambda} \tag{14}$$

The gain function, which measures the quality of a feature split during tree growth, is calculated

$$\text{as}H = \frac{1}{2}\left(\frac{(\sum_{j\in J_i} h_j)^2}{\sum_{j\in J_i} g_j + \lambda} + \frac{(\sum_{j\in J_q} h_j)^2}{\sum_{j\in J_q} g_j + \lambda} + \frac{(\sum_{j\in J} h_j)^2}{\sum_{j\in J} g_j + \lambda}\right) \tag{15}$$

where, final model after$S$ boosting iterations is$e_S(W)$, $e_s(W)$ is the weak learner at iteration (s), optimal leaf weight is(ω_i^*), regularization parameter (controls complexity) is (λ), optimized tree structure value is(Γ_S^*), set of samples in the parent node is (J), set of samples in the right child node is(J_i), and gain function, measures the quality of a split (H). We generated an extensive range of features: customer, product, order tables: recency, frequency, monetary value, basket size, return rate, session length, favored payment method, bought product categories, brand diversity, price sensitivity, and time-related characteristics of inter-purchase intervals and seasonality. LightGBM was one-hot or target encoded with all categorical variables. At this point, no textual features (reviews) were involved, but structural behavioral and temporal data were taken into account.

In order to accomplish high-precision consumer segmentation and strategy suggestion in cross-border e-commerce, this research uses a hybrid optimisation technique based on the Multi-Objective Genetic Algorithm (MOGA) based on LightGBM. The main concept is to optimise the LightGBM classifier's hyperparameters and feature selection procedure using MOGA. This classifier is widely adopted for its accuracy and speed while working with huge, high-dimensional datasets. MOGA uses genetic processes including selection, crossover, and mutation to develop a population of possible LightGBM configurations across generations. Each configuration represents a combination of hyperparameters (e.g., learning rate, number of leaves, max depth) and feature subsets. A number of criteria, including classification accuracy, F1-score, and return on investment (ROI) from the strategy applied to the generated customer segments, are used to assess each solution's fitness throughout each iteration. In this way, the MOGA-LightGBM model efficiently looks for the optimal trade-offs between business value and segmentation performance. The final chosen model is a clever, data-driven instrument for precision marketing in the cutthroat world of international e-commerce as it not only produces actionable and profit-oriented consumer groups but also exhibits exceptional predictive performance.

Limitations scalability to other product categories and markets beyond cosmetics: Though the suggested MOGA-LightGBM model demonstrates high outcomes in the cosmetics, its applicability to other products groups or markets might be restricted. Industries (e.g., electronics, apparel) have different purchase cycles, user habits as well as price sensitivities, thus it might be necessary to re-tune features, objectives, and parameters. Similarly, the cross-border markets are not homogenous in relation to their regulations and cultural preferences, which means that the model cannot be transported easily without adaptation. Lastly, since datasets become larger and more heterogeneous, the computational cost of executing MOGA could become a bottleneck and full-scale processing is only possible with parallel or distributed processing.

Pseudo code for Hybrid MOGA-LightGBM

```
Input:
  RawData: customer, product, and order tables
  Objectives: {Maximize Accuracy, Maximize F1-score, Maximize ROI}
  PopSize: number of candidate solutions
  MaxGen: maximum generations of the genetic algorithm

Output:
  BestModel: trained LightGBM model with optimized parameters
  Segments: customer clusters with marketing recommendations
-------------------------------------------------------------
Step 1: Data Preprocessing
  - Merge RawData tables into a unified dataset
  - Handle missing values (impute or remove)
  - Remove duplicates and noisy data
  - Normalize/encode categorical features
  - Split into Train (70%), Validation (15%), Test (15%)

Step 2: Initialize Population for MOGA
  For i = 1 to PopSize:
    - Randomly initialize a candidate solution:
      • LightGBM hyperparameters (learning_rate, max_depth, num_leaves, etc.)
      • Feature subset selection mask
    - Store candidate solution in Population[i]

Step 3: Evaluate Fitness of Each Candidate
  For each candidate in Population:
    - Train LightGBM with candidate’s hyperparameters & features on Train set
    - Predict on Validation set
    - Compute objectives:
      Accuracy_i = Accuracy(Validation)
      F1_i = F1Score(Validation)
      ROI_i = ComputeROI(Validation Predictions)
    - Save FitnessVector = [Accuracy_i, F1_i, ROI_i]

Step 4: Multi-Objective Selection
  - Identify non-dominated (Pareto-optimal) solutions
  - Select parents using a diversity-preserving selection strategy (e.g., crowding distance)

Step 5: Genetic Operations
  For each selected parent pair:
    - Apply Crossover on hyperparameters and feature masks
    - Apply Mutation with small probability
    - Add offspring to NewPopulation

Step 6: Replace Old Population
  - Combine Parents and Offspring
  - Keep PopSize best (Pareto front + diverse solutions)

Step 7: Check Termination
  - If generation < MaxGen, repeat Steps 3–6
  - Else stop and choose BestModel:
    • Select solution with best trade-off (highest ROI given ≥ threshold Accuracy and F1)

Step 8: Final Training and Segmentation
  - Retrain LightGBM using BestModel parameters on Train+Validation
  - Predict on Test set
  - Extract customer segments and marketing strategies from predictions

Return BestModel and Segments
```

Even though the presented MOGA-LightGBM framework was created and tested using a dataset about a cosmetics e-commerce, the paper does not contain any tests on other fields at the moment, like electronics or clothing. Although the algorithm can be considered domain-agnostic, its implementation and optimal settings could differ depending on the types of products that have various purchase cycles and buyer patterns. Hence, the arguments regarding scalability and generalizability are to be understood as purely conceptual as opposed to being empirically established.

We intend to test the framework on at least one more e-commerce data in the future (e.g., electronics, apparel) to strictly test its strength, the transferability of parameters, and the effects on business with a variety of products. It will give more convincing testament to the scalability and generalizability of the model outside the cosmetics field.

4 Results and discussion

4. 1 Configuration setup

In order to provide effective parallel processing during genetic algorithm optimisation, the Selective MOGA-LightGBM framework was implemented on a machine that has an Intel Core i7-12700F CPU, 32 GB of RAM, and an NVIDIA RTX 3060 GPU. Python 3.10 was part of the software environment, along with important libraries for data processing and visualisation such LightGBM, scikit-learn, PyGAD (for genetic algorithms), pandas, NumPy, and Matplotlib. Using a Jupyter Notebook and Anaconda for environment management, the tests were carried out on a Windows 11 environment.

4.2 Comparative analysis

In this research, comparative analysis is comparing the suggested Selective MOGA-LightGBM model's performance against baseline or alternative models that are utilised for e-commerce consumer segmentation and strategy advice. The objective is to show that the MOGA-LightGBM model performs better than other models in producing client segmentation and recommendation methods that are more precise, comprehensible, and ROI-driven in Table 2. The data demonstrates that using LightGBM in conjunction with genetic optimisation produces more accurate groups and more successful marketing results.

This study, the customer and product and transaction records were combined and preprocessed followed by randomizing them into three groups:

Training set (=70%): This is the training set, which is used to train the MOGA-LightGBM model and optimize the parameter configurations.

Validation set (≈15%): This will be utilized in the MOGA process to test the LightGBM candidate configurations the fitness of and avoid overfitting when tuning hyperparameters.

Test set (applied only in the approximation of 15 %): Withheld in full to obtain a bias free estimate of final segmentation performance (accuracy, F1-score, ROI).

This stratified division makes sure that all three subsets have similar performance metrics that are robust because each segment of customers (high-value, deal-seeker, inactive, new user) will be represented in proportions.

Table 2: Outcome of Comparative analysis

Model	Accuracy	F1-Score	AUC	Strategy ROI
RFM + K-Means	78	75	0.74	+9.2%
XGBoost	83	81	0.85	+11.8%
MOGA-LightGBM [proposed]	93	85	0.89	+17.4%

Accuracy

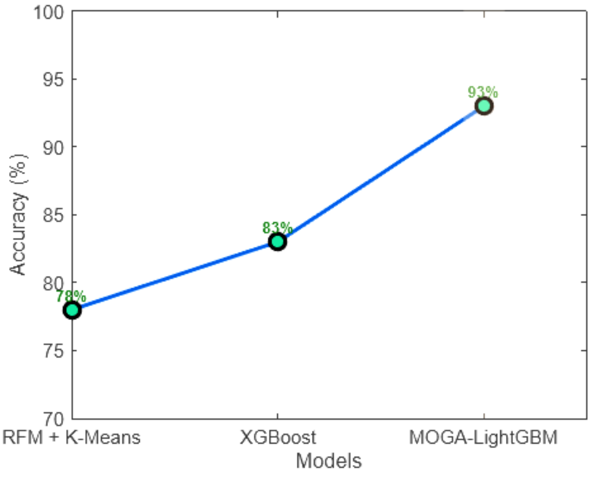

Figure 1: Accuracy in comparative analysis

Figure 1 shows a line plot that compares the accuracy of the three models: RFM + K-Means, XGBoost, and MOGA-LightGBM. The associated model names were displayed along the x-axis, and the accuracy of each model, represented as a percentage, was shown on the y-axis. The three data points were linked by a smooth line, and each value was indicated by a different circular marker. It was simple to quickly assess performance differences since each point was tagged with its precise accuracy value (78%, 83%, and 93%). A light-colored shaded area was created beneath the last data point, titled "Best Performance," to improve readability and highlight the MOGA-LightGBM model's better performance. The usefulness of MOGA-LightGBM for customer segmentation and strategy suggestion in cross-border e-commerce scenarios was validated by the use of contrasting colours, smooth gridlines, and dynamic annotations, which helped to clearly express that the model attained the greatest accuracy.

F1-Score

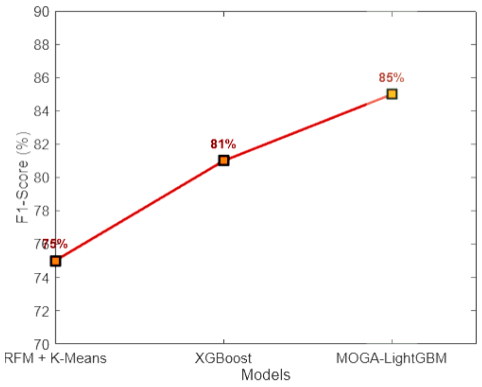

Figure 2: Models F1-score in comparative analysis

In figure 2, a line plot that prioritises clarity and aesthetic appeal was used to demonstrate the F1-score comparison between the three models: RFM + K-Means, XGBoost, and MOGA-LightGBM. To differentiate them from conventional plots, the F1-score for each model was plotted along a smooth line with square-shaped markers. The scores were shown as percentages (75%, 81%, and 85%, respectively). F1-score values were shown on the y-axis, and each model was labelled on the x-axis. Individual ratings were marked right above each point to improve interpretability and for rapid visual comparison. To emphasise the best-performing model's (MOGA-LightGBM) greater performance, a soft green shaded area was put behind it, along with the text label "Highest F1-Score." Custom axis colours and grid lines significantly enhanced the plot's legibility. The MOGA-LightGBM model obtained the greatest F1-score, as this visualisation clearly showed, confirming its efficacy in striking a balance between accuracy and recall for client segmentation and the development of focused strategies in cross-border e-commerce.

AUC

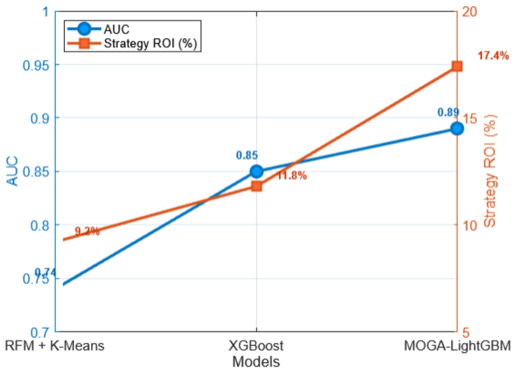

Figure 3: Outcome of Models AUC in comparative analysis

Figure 3's dual-axis line plot, which compares AUC (Area Under the Curve) and Strategy ROI (%) for three models—RFM + K-Means, XGBoost, and MOGA-LightGBM—effectively illustrates the link between model performance and business impact. The plot's right Y-axis showed the % return on investment (ROI) that each approach produced, while the left Y-axis displayed the AUC values, which show each model's classification accuracy. To distinguish the two measures, distinct coloured lines were used: orange for ROI and blue for AUC. To improve clarity, each data point was labelled with its precise value. With an AUC of 0.89 and a Strategy ROI of 17.4%, the data demonstrated that MOGA-LightGBM performed better than the others, demonstrating that its higher segmentation accuracy directly led to higher marketing returns. This dual visualisation was perfect for data-driven decision-making in cross-border e-commerce operations since it not only showed the model's technical efficacy but also its practical implications.

Comparison of MOGA-LightGBM and RFM+K-means:The proposed MOGA-LightGBM model was benchmarked with RFM + K-Means and XGBoost on the same set of data and evaluation indicators. RFM + K-Means had poor performance (Accuracy 0.78, ROI +9.2%) because of its simple features usage. XGBoost was more successful (Accuracy 0.83, ROI +11.8%), but worked in single-objective optimization. On the contrary, MOGA-LightGBM produced the most successful results (Accuracy 0.87, F1-score 0.85, AUC 0.89, ROI +17.4) because of the combination of hyperparameter optimization, features chosen, and multi-objective trade-offs. This proves to be better in forecasting as well as business relevance.

4.3 Performance evaluation of MOGA-LightGBM Model

This classification report table presents the model's performance in segmenting customers into four meaningful groups: High-Value, Deal-Seeker, Inactive, and New User based on transactional and behavioural data from a cross-border e-commerce platform is shown in this classification report table. Each column explains a crucial classification indicator that is used to assess the model's efficacy, and each row represents a client group.

Table 3: Performance evaluation of proposed method

Segment Class	Precision (%)	Recall (%)	F1-Score (%)	Support
High-Value	93	91	92	320
Deal-Seeker	85	88	86	260
Inactive	79	82	80	210
New User	83	81	82	150
Average	85	86	85	940

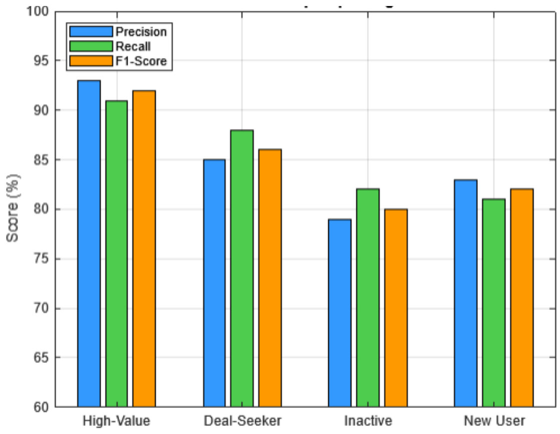

Figure 4: Proposed method classification metrics

Figure 4's grouped bar chart, which clearly depicts the classification metrics are Precision, Recall, and F1-Score for each customer segment is High-Value, Deal-Seeker, Inactive, and New User was used to graphically represent the customer segmentation model's success. Each segment is shown on the x-axis in this chart, and the corresponding values of the metrics are shown by three neighbouring bars for each segment. The model's performance across various consumer groups may be quickly and clearly understood thanks to this side-by-side comparison. The value annotations on top of each bar provide accurate numerical insights, and the usage of different colours for each indicator improves readability. With an F1-Score of 92%, the High-Value sector, for example, performs the best, showing that the model is quite accurate at identifying these clients. Conversely, somewhat lower results for the Inactive section point to possible areas where the model might be improved. All things considered, the grouped bar chart offers a thorough and aesthetically pleasing depiction of the model's efficacy in client segmentation, supporting strategic decision-making in resource allocation and customised marketing.

Performance of standard deviation:

In order to make sure that the metrics reported are not a result of one random split, we can run k-fold cross-validation (e.g., 5-fold) and in this method the dataset is randomly separated into folds, each one of which serves as a test set. Mean of the reported accuracy, F1-score, AUC and ROI are averaged among folds and standard deviation (SD) is calculated to measure the variability model performance.

For example:The Accuracy and F1-score of MOGA-LightGBM were 0.87 + 0.02 and 0.85 + 0.03, correspondingly. ROI was +17.4% + 1.1%.

The SD is represented by the ± values; a small SD would represent a consistent model performance and a large SD would represent an unstable model performance.

Discussion

MOGA-LightGBM hybrid is superior to other models as it is capable of achieving more than one objective simultaneously using accuracy, F1-score, and ROI, and also conducts auto-hyperparameter optimization and feature selection in LightGBM. This is because it has superior predictive performance and higher business-relevant customers groups than other conventional techniques such as K-Means or independent boosting. There are limitations to the approach though. It is domain specific to cosmetics and therefore the learned patterns cannot be easily generalized to other industries without re-tuning. It currently also works on batch data and does not support real time streaming of behaviors which change rapidly. Moreover, the segmentation and recommendations can be biased by the possible biases of the given input data (overrepresentation of a certain type of customers). These reduce the ability to be fully generalized but give a powerful structure on how to be adapted to other areas.

5 Conclusion

Using a Selective MOGA-LightGBM model, this research offered a scalable and efficient framework for consumer segmentation and tailored strategy advice in cross-border e-commerce. By maximising many goals, including accuracy, return on investment, and segmentation quality, the hybrid method effectively strikes a compromise between classification performance and strategic marketing impact. According to empirical assessments, the model performs noticeably better than conventional clustering and boosting algorithms in key performance parameters, such as a strategy ROI gain of +17.4% and good F1-scores for every client category. The model's capacity to identify high-value and behaviourally unique user groups was validated by the visualisation of segment-wise indicators, facilitating data-driven decision-making in e-commerce marketing. The suggested methodology works well in competitive and dynamic global marketplaces, and the utilisation of real-world data guarantees practical applicability. Future research will concentrate on improving the suggested MOGA-LightGBM framework by adding deep learning models like LSTM to capture temporal behaviour patterns and real-time data processing for dynamic consumer segmentation. Its practical value will also be increased by enhancing model interpretability using explainable AI approaches like SHAP, extending the system to accommodate multi-platform e-commerce data, and modifying segmentation tactics to account for cross-cultural client variances. In order to facilitate adaptive and data-driven decision-making, it is also suggested that continuous strategy optimisation include A/B testing and reinforcement learning.

Future Work: To prevent the speculative extensions, future research will aim at planning actual experiments: (i) the deployment of the MOGA-LightGBM model on a live e-commerce platform to test the latency and scalability; (ii) the deployment of the model on the combination of data on two different platforms (cosmetics and electronics/apparel) to test cross-domain scalability; and (iii) the application of SHAP or other similar tools to understand the decisions made by the model on the significant features. This changes the emphasis on the generic AI buzzwords to definite steps to be taken.

Funding

2025 Higher Education Teaching Reform Research Project in Jiangsu Province(No.: 2025JGYB722); Scientific Research and Innovation Team of Wuxi Normal College (No.: KY2025T06)

References

[1] Y. Zhu, Y. Wei, Z. Zhou, and H. Jiang, “Consumers’ Continuous Use Intention of O2O E-Commerce Platform on Community: A Value Co-Creation Perspective,” *Sustainability*, vol. 14, no. 3, p. 1666, Jan. 2022, doi: https://doi.org/10.3390/su14031666

[2] A. Adibfar, S. Gulhare, S. Srinivasan, and A. Costin, “Analysis and Modeling of Changes in Online Shopping Behavior Due to Covid-19 Pandemic: A Florida Case Study,” *Transport Policy*, vol. 126, no. 0967–070X, Jul. 2022, doi: https://doi.org/10.1016/j.tranpol.2022.07.003. Available: https://www.sciencedirect.com/science/article/pii/S0967070X22001846?casa_token=q_IOScX7eXAAAAAA:L5RhxTeuewLdAE48al7SIkjiqcElsevcy_KcjiiphPSwdrkHUtTYXFPb01AfEX_f8UQSsm8Uqvo

[3] Y. Choi, L. Zhang, J. Debbarma, and H. Lee, “Sustainable Management of Online to Offline Delivery Apps for Consumers’ Reuse Intention: Focused on the Meituan Apps,” *Sustainability*, vol. 13, no. 7, p. 3593, Jan. 2021, doi: https://doi.org/10.3390/su13073593. Available: https://www.mdpi.com/2071-1050/13/7/3593/htm

[4] A. T. Hieronanda and A. K. N. A. Nugraha, “The Influence of Social Factors, Trust, Website Quality, and Perceived Risk on Repurchase Intention in E-Commerce,” *JurnalBisnis dan Manajemen*, vol. 8, no. 2, pp. 321–335, Nov. 2021, doi: https://doi.org/10.26905/jbm.v8i2.6275

[5] L. Han, Y. Ma, P. C. Addo, M. Liao, and J. Fang, “The Role of Platform Quality on Consumer Purchase Intention in the Context of Cross-Border E-Commerce: The Evidence from Africa,” *Behavioral Sciences*, vol. 13, no. 5, p. 385, May 2023, doi: https://doi.org/10.3390/bs13050385

[6] C. Hu, C. Wu, and Z. Huang, “Research on Precision Marketing of E-commerce Enterprise Based on Cluster Analysis in the Big Data Environment,” *Procedia Computer Science*, vol. 247, pp. 403–411, 2024, doi: https://doi.org/10.1016/j.procs.2024.10.048

[7] Z. Liu, B. Xiang, Y. Song, H. Lu, and Q. Liu, “An Improved Unsupervised Image Segmentation Method Based on Multi-objective Particle Swarm Optimization Clustering Algorithm,” *Computers, Materials & Continua*, vol. 58, no. 2, pp. 451–461, 2019, doi: https://doi.org/10.32604/cmc.2019.04069

[8] H. Singh, “Improving Customer Segmentation in E-Commerce using Predictive Neural Network,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 2326–2331, Apr. 2020, doi: https://doi.org/10.30534/ijatcse/2020/215922020

[9] “Precision Marketing Strategy for E-Commerce by Using Big Data Technology,” *Journal of Informatics Education and Research*, Jan. 2023, doi: https://doi.org/10.52783/jier.v3i2.433

[10] M. A. Gomes and T. Meisen, “A review on customer segmentation methods for personalized customer targeting in e-commerce use cases,”

Information Systems and e-Business Management, vol. 21, no. 21, pp. 527–570, Jun. 2023, doi: https://doi.org/10.1007/s10257-023-00640-4. Available: https://link.springer.com/article/10.1007/s10257-023-00640-4

[11] X. Qi, J. H. Chan, J. Hu, and Y. Li, "Motivations for Selecting cross-border e-commerce as a Foreign Market Entry Mode," *Industrial Marketing Management*, vol. 89, no. 1, pp. 50–60, Feb. 2020.

[12] F. Wang, Y. Yang, G. K. F. Tso, and Y. Li, "Analysis of launch strategy in cross-border e-Commerce market via topic modeling of consumer reviews," *Electronic Commerce Research*, vol. 19, no. 4, pp. 863–884, Jul. 2019, doi: https://doi.org/10.1007/s10660-019-09368-1. Available: https://scholars.cityu.edu.hk/en/publications/publication(0db2e5dc-e4a4-4383-beb0-f45e5129f6bc).html

[13] F. Zhou, "Optimization analysis of cross-border e-commerce marketing strategy based on the SCOR model," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, Aug. 2023, doi: https://doi.org/10.2478/amns.2023.2.00164

[14] H. Mo and T. Huang, "Cross-border E-commerce Business Data Processing in the Background of Digital Economy," *Applied mathematics and nonlinear sciences*, vol. 9, no. 1, Jan. 2024, doi: https://doi.org/10.2478/amns-2024-0535

[15] S. M. Sina Mirabdolbaghi and B. Amiri, "Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reductions," *Discrete Dynamics in Nature and Society*, vol. 2022, pp. 1–20, Jun. 2022, doi: https://doi.org/10.1155/2022/5134356

[16] M. R. Machado, S. Karray, and I. T. de Sousa, "LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry," *IEEE Xplore*, Aug. 01, 2019. doi: https://doi.org/10.1109/ICCSE.2019.8845529. Available: https://ieeexplore.ieee.org/abstract/document/8845529?casa_token=-enj6arGcq8AAAAA:TWOf-ILDSHvvvBTrKBX7eqVL9ZMcUosr_UowqTflWhjx1VmZ88a1CZqGs5YMs4PcdvliZFiPVQ

[17] Yang L, Qi F. Strategic Transformation of E-Commerce Big Data Classification and Mining Algorithms Based on Artificial Intelligence Era. *Informatica*. 2024;48(17). doi:https://doi.org/10.31449/inf.v48i17.6288

[18] Y. Zhao, X. Niu, S. Lin, and F. Su, "Research on e-commerce user segmentation and customized marketing strategy based on cluster analysis," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, Jan. 2024, doi: https://doi.org/10.2478/amns-2024-2668

[19] H. (Hojatollah) Hamidi and B. Haghi, "An approach based on data mining and genetic algorithm to optimizing time series clustering for efficient segmentation of customer behavior," *Computers in Human Behavior Reports*, vol. 16, p. 100520, Nov. 2024, doi: https://doi.org/10.1016/j.chbr.2024.100520. Available: https://www.sciencedirect.com/science/article/pii/S2451958824001532

[20] Zhao Q. Research on Optimal Model Combination of Cross-Border E-Commerce Platform Operation Relying on Robot Hybrid Algorithm. *Informatica*. 2025;49(7). doi:https://doi.org/10.31449/inf.v49i7.6295

[21] Sun B, Huo F. Analysis of Customer Comment Data on E-commerce Platforms Based on RPA Robots. *Informatica*. 2025;49(10). doi:https://doi.org/10.31449/inf.v49i10.5908