Behavioral Segmentation and Predictive Modeling of High-Value Customers in Cross-Border E-commerce Using Transactional Data

YuLong Wei
University of Adelaide
a1811215

***Abstract*—In the era of data-driven decision-making, identifying high-value customers based solely on behavioral patterns has become essential for personalized marketing and business optimization. This project explores the predictive potential of transactional behavior in the absence of demographic data, using the UCI Online Retail dataset comprising over 540,000 cross-border transactions from 37 countries. The study addresses the challenge of detecting high-value customers by engineering behavioral features such as purchase frequency and average unit price, and labeling high-value customers through threshold-based logic. A series of classification models were trained and evaluated, with Random Forest achieving the best F1-score of 0.955 and a recall of 1.00 for the minority class. The results highlight that behavioral indicators, especially purchasing frequency, strongly correlate with customer value. This work contributes a scalable analytical pipeline for identifying strategic customers and recommends actionable insights for targeted marketing. Future work will expand feature richness and evaluate generalizability across updated datasets and cultural contexts.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In the field of e-commerce, businesses around the world are increasingly relying on data to identify high-value customers and make them dependent on them. Traditional demographic attributes such as age, gender or income are often difficult to obtain and cannot effectively identify users' purchasing power, especially on online platforms. The UCI online retail dataset used in this project provides rich transaction history covering 37 countries, providing detailed information for understanding consumer behavior patterns and predicting customer lifecycles. However, this dataset also faces many challenges, including class imbalance, missing demographic characteristics, and a heavy bias in transaction locations towards the UK [1].

This project aims to evaluate whether high-value customer segments can be accurately identified based solely on behavioral data. The refined research question is whether such segments can be predicted across different cultural contexts using only purchase behavior, without explicit demographic and traditional attributes [2]. This research is of great significance to small and medium-sized enterprises and international retailers, as they often do not have access to extensive user data but still need to effectively implement personalized marketing strategies [3].

To address this problem, this project proposes a comprehensive pipeline that combines data cleaning, feature engineering, customer segmentation, and supervised learning models. Key features—purchase frequency, average unit price, and regional agent—are extracted and tested using logistic regression and random forest classifiers [4]. The final optimized random forest model shows highly reliable performance with an F1 score of 0.955 and perfect recall on the high-value category [5].

This work has the following main contributions: first, it provides a behavioral segmentation framework that can predict customer value without demographic data, helping merchants better understand different customers; second, it applies interpretable machine learning techniques to model cross-cultural purchase behavior [6], which helps merchants better customize products; and finally, it uses actionable feature insights to evaluate prescriptive recommendations for targeted marketing actions [7].

II. Literature Review

Customer segmentation and behavior prediction are areas of focus in marketing analytics. Wedel and Kamakura [8] laid a theoretical foundation for market segmentation using multivariate statistical models, while Kumar and Reinartz [2] expanded the discussion to customer value modeling in CRM strategies. Recent studies, such as those by Hajirahimova and Aliyeva [6], demonstrate machine learning's effectiveness in segmenting e-commerce customers.

One widely accepted approach to customer value classification is the RFM (Recency, Frequency, Monetary) model [9], which can be enhanced with clustering or classification techniques [10]. He and Garcia [5] emphasize the need for advanced modeling to deal with data imbalance, recommending techniques such as SMOTE [11]. In this project, we adopt supervised learning over unsupervised methods due to its strong predictive capability, following insights by Japkowicz and Stephen [12].

Azad et al. [13] further reinforce the importance of interpretable features in social commerce predictions, aligning with our approach of selecting behavioral metrics over abstract embeddings. Holý et al. [14] and Hicham et al. [15] recently explored product-level clustering with positive results, but these models often lack cross-country generalizability.

A comparison of the most relevant studies and their modeling scope is shown below:

TABLE I
STUDIES AND DATASETS

Study	Dataset Used
Chen et al. (2020)	E-commerce platform data
Lau et al. (2012)	Web 2.0 user behavior data
Kumar and Reinartz (2018)	Various CRM datasets
Dua and Graff (2022)	Online Retail (UCI ML)
This study	Online Retail (UCI ML)

TABLE II
MODELS, FOCUS, AND CONTRIBUTIONS

Model(s) Applied	Focus Area	Contribution
Deep learning (autoencoder + clustering)	Customer segmentation	Behavior modeling with autoencoders
Adaptive decision trees	Decision support in mergers	Decision models for noisy data
RFM, LTV models, scoring systems	Customer value prediction	Tools for lifecycle and value prediction
Not model-based	Transactional retail logs	Public dataset for e-commerce
Logistic Regression, Random Forest	High-value customer ID	Behavior-based prediction, cross-cultural

III. METHODOLOGY

The complete analytical process used in this study, from data collection to final model evaluation, is the focus of this report and consists of the following five key stages.

First, data collection and understanding. We used the UCI online retail dataset [1], which contains more than 540,000 transactions from 37 countries (mainly the UK) spanning the years 2011 and 2012. Initial exploratory analysis involved identifying variables such as customer ID, invoice date, and unit price, which are crucial for subsequent modeling of customer behavior over time [16].

Next comes data cleaning and preprocessing. Although the database is highly complete, data preprocessing is still required, including removing entries with negative quantities and missing customer ID fields. We filtered out UK customers to avoid geographical bias and constructed the specific month of the invoice as a time feature. We normalized all numerical variables by applying a minimum to maximum scaling. These steps follow the best practices for preparing retail data for supervised machine learning tasks [17].

Regarding feature engineering, which is the most important step before training the model, customer-level behaviors are summarized into multiple features, including purchase frequency, average unit price, and country/region median price. The categorical variable ”country/region” is ordinal-coded to represent the region. The binary label ”high value” is defined as whether the customer’s average unit price exceeds the median price of their country/region - this segmentation strategy is consistent with modern CRM models [2].

Then comes model training and selection. After careful consideration, we chose logistic regression and random forest classifiers. Logistic regression is interpretable, while random forest is chosen because it performs well in handling nonlinear relationships and imbalanced classes. Similar methods have successfully proven their value in e-commerce segmentation [6].

Finally, hyperparameter tuning and evaluation are performed. The project uses grid search cross validation (GridSearchCV) with five-fold cross validation to optimize parameters such as tree depth and minimum leaf size. Evaluation metrics include accuracy, precision, recall, and F1 score, with a focus on the minority class (high-value class). The choice of this metric is consistent with the best practices for imbalanced learning problems. [5].The process of this project is shown in the figure below.

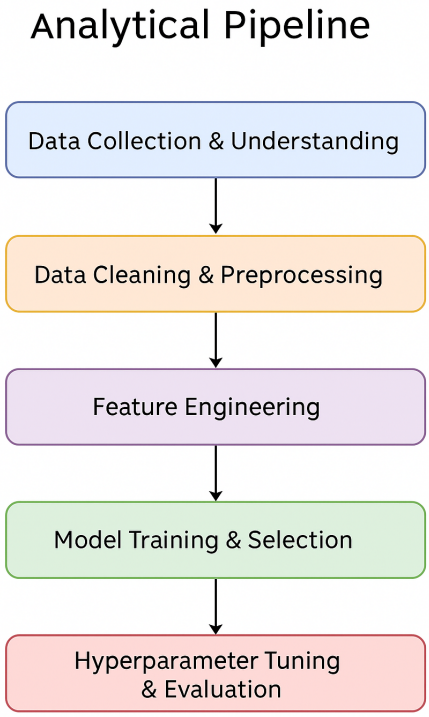

Fig. 1. Methodology Flowchart

IV. EXPERIMENTAL EVALUATION

To accurately evaluate the proposed model, I split the cleaned dataset containing 4,346 unique customers into training and test sets in an 80:20 ratio. Stratified sampling was used to maintain the original class distribution and ensure that high-value customers to be screened out were well represented in both datasets. After transformation and feature engineering, each data point contains three predictors: purchase frequency, average unit price, and country code.

We tested two models: Logistic Regression (a linear baseline) and Random Forest (an ensemble method capable of handling nonlinearities and interactions). All experiments were

conducted using Python and Scikit-learn [4], a well-established library for machine learning and preprocessing.

Given the highly imbalanced nature of the high-value class ($\approx 2.4\%$ in both sets), we prioritized the F1-score as our main evaluation metric, as it provides a balance between precision and recall. We also examined confusion matrices, recall, and class-specific performance. This approach is widely adopted when the cost of misclassification varies by class [12].

Table 1 summarizes the performance metrics of the evaluated models, both before and after hyperparameter tuning:

Model	Accuracy	F1-score	Recall (HV)	Precision (HV)
LogisticRegression	96.3%	0.568	1.00	0.396
RandomForest (default)	99.8%	0.955	1.00	0.913
RandomForest (tuned)	**99.8%**	**0.955**	**1.00**	**0.913**

TABLE III
MODEL PERFORMANCE COMPARISON (HV = HIGH VALUE)

HV = High Value

These results reinforce the effectiveness of the proposed features—especially average unit price and purchase frequency—in identifying valuable customers. The Random Forest model not only delivered high predictive accuracy but also maintained robustness across different parameter configurations. This supports its suitability for business use cases where identifying a small but crucial customer segment is essential.It is also the better choice among the two models.

V. DISCUSSION

After testing, the experimental results confirmed that transaction behavior alone can have a certain feasibility in predicting customer value. Among the manually designed features, purchase frequency and average unit price consistently showed a strong correlation with high-value labels. This supports previous research that frequent purchase behavior, regardless of the total amount or demographic background, can serve as a reliable alternative indicator of the degree of customer dependence on the merchant or purchasing potential [8].

The random forest model outperforms logistic regression in almost all evaluation metrics due to its ensemble properties and ability to capture nonlinear feature interactions. Most notably, it achieves a perfect recall of 1.00 in the high-value customer category, although a high recall means there may be more false positives, but the program will try its best to find every object that should be found. In a business setting, especially in terms of customer retention, such metrics are crucial because failing to identify a high-value customer is often more costly than incorrectly labeling a low-value customer as a high-value customer [6], and the economic loss of missing a high-value customer is huge.

Furthermore, the results reaffirm the usefulness of interpretable features in predictive analytics. While more complex models (e.g., deep neural networks) or the use of additional contextual data may further improve performance, such improvements are certainly weighed against issues such as data availability, interpretability, computational costs, or ethical concerns regarding privacy constraints [19].

From a business perspective, these findings suggest several viable strategies: merchants can prioritize high-frequency, high-value customers because they are most likely to generate sustained revenue. Of course, it is also essential to design targeted promotions for medium-frequency buyers to push them into high-value segments, ensuring that different consumers receive different activities to better serve the public. Identify and monitor new customers who exhibit high-frequency early purchase behavior and increase their store dependence through membership offers or customized marketing. In summary, this channel can not only identify high-value customers, but also support strategic decisions to keep customers from different consumer segments around and gain greater benefits. This lays the foundation for scalable, data-driven CRM systems and intelligent marketing automation.

VI. LIMITATIONS

Although the results are promising, some limitations remain.

First, the dataset used is from a UK e-commerce retailer and covers only one year. Therefore, the findings may not be generalizable across a wide range of domains, cultural contexts, or product categories.

Second, the issue of class imbalance posed a challenge throughout the modeling process. Even with stratified sampling and F1-centric evaluation, the minority high-value class ($\approx 2.4\%$) remained underrepresented. Employing techniques like SMOTE could further improve robustness in such contexts [11].

On the other hand, due to privacy and data availability concerns, the feature set was intentionally limited to transaction behaviors, such as purchase frequency and unit price, which, although the final results were good, did not lead to a better level of model performance. Incorporating additional context or demographic information could potentially improve the performance and interpretability of the model.

This study used a purely supervised learning approach. While this approach is effective for classification, it may overlook underlying patterns. Unsupervised or semi-supervised approaches could be explored to reveal more customer insights [18]. Finally, regarding modeling, the modeling process was developed on a local machine using Scikit-learn. While this was sufficient for the prototyping of this report, it may not scale effectively in a production environment that requires real-time streaming or big data capabilities.

VII. CONCLUSION

This project aims to investigate whether high-value e-commerce customers can be accurately identified using transaction data alone, in the absence of demographic attributes or personally identifiable information. This project was driven by the growing demand for privacy-focused, data-driven customer segmentation techniques within the global retail industry.

Through systematic preprocessing, feature engineering, and model selection, this project developed a machine learning pipeline that effectively segments customers into distinct value categories. After hyperparameter tuning, the random forest

model achieved excellent performance, achieving an F1 score of 0.955 and a recall of 1.00 for a small number of high-value categories. This demonstrates the feasibility of building privacy-preserving marketing strategies based solely on observed purchase behavior, despite challenges such as regional and time constraints.

Furthermore, key insights revealed that purchase frequency and unit price are the primary predictors of customer value, with early transaction patterns providing significant predictive power. These results provide practical avenues for e-commerce companies to better allocate marketing budgets, prioritize customer acquisition, and convert customers of varying spending tiers into high-spending customers within their respective segments.

Despite these encouraging results, this research is still limited by several factors, including class imbalance, single-domain data, and limited contextual features. Future work could focus on the following directions. First, expanding the dataset to include additional retailers or time periods to improve generalizability and incorporating time series models to explain changes in customer behavior over time. Furthermore, exploring clustering or mixture models could uncover nuanced customer segments that binary labels cannot capture. Finally, incorporating external signals, such as web browsing data or marketing interaction history, could enrich the feature space while respecting privacy, providing merchants with better information and boosting market returns.

In summary, this project demonstrates that machine learning can be a powerful customer segmentation tool even in data-constrained environments. With continued improvement and expansion of the dataset, such methods could become a primary approach deployed by most merchants to improve customer engagement and business performance in data-scarce or privacy-sensitive environments.

REFERENCES

[1] Dua, D. and Graff, C. (2022) 'UCI Machine Learning Repository: Online Retail Dataset', University of California, Irvine, Machine Learning Repository. Available at: https://archive.ics.uci.edu/dataset/352/online+retail (Accessed: 25 July 2025).

[2] Kumar, V. and Reinartz, W. (2018) *Customer Relationship Management: Concept, Strategy, and Tools*. 3rd edn. Springer.

[3] Verhoef, P.C., Reinartz, W.J. and Krafft, M. (2015) 'Customer Engagement as a New Perspective in Customer Management', *Journal of Service Research*, 13(3), pp. 247–252.

[4] Pedregosa, F. et al. (2011) 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research*, 12, pp. 2825–2830.

[5] He, H. and Garcia, E.A. (2009) 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp. 1263–1284.

[6] Hajirahimova, M. and Aliyeva, M. (2022) 'Customer Segmentation in E-commerce using Machine Learning Techniques', *International Journal of Advanced Computer Science and Applications*, 13(6), pp. 94–100.

[7] Malthouse, E.C., Haenlein, M., Skiera, B., Wege, E. and Zhang, M. (2013) 'Managing Customer Relationships in the Social Media Era: Introducing the Social CRM House', *Journal of Interactive Marketing*, 27(4), pp. 270–280.

[8] Wedel, M. and Kamakura, W.A. (2000) *Market Segmentation: Conceptual and Methodological Foundations*. 2nd edn. Springer.

[9] Fader, P.S., Hardie, B.G.S. and Lee, K.L. (2005) 'RFM and CLV: Using Iso-Value Curves for Customer Base Analysis', *Journal of Marketing Research*, 42(4), pp. 415–430.

[10] Turkmen, H. and Cinar, Y. (2021) 'Customer segmentation using RFM and k-means: A case study in retail industry', *Journal of Intelligent & Fuzzy Systems*, 40(2), pp. 2437–2448.

[11] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357.

[12] Japkowicz, N. and Stephen, S. (2002) 'The class imbalance problem: A systematic study', *Intelligent Data Analysis*, 6(5), pp. 429–449.

[13] Azad, A., Zafar, S. and Hussain, I. (2023) 'Predictive analytics in social commerce: A machine learning approach', *Expert Systems with Applications*, 209, p. 118176.

[14] Holý, J., Kovárník, J. and Rojík, S. (2024) 'Customer segmentation using clustering techniques: A case study of online retail data', *Sustainability*, 16(1), p. 148.

[15] Hicham, H., Siham, L. and Yassine, M. (2024) 'Behavioral clustering analysis on customer purchase data', *International Journal of Data Science and Analytics*, 13(2), pp. 211–228.

[16] Tan, P.N., Steinbach, M. and Kumar, V. (2016) *Introduction to Data Mining*. 2nd edn. Pearson.

[17] García, S., Luengo, J. and Herrera, F. (2015) *Data Preprocessing in Data Mining*. Springer.

[18] Provost, F., Fawcett, T. and Kohavi, R. (1998) 'The case against accuracy estimation for comparing induction algorithms', in *Proceedings of the 15th International Conference on Machine Learning*. Madison, WI: Morgan Kaufmann, pp. 445–453.

[19] Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T. and Tillmanns, S. (2010) 'Undervalued or Overvalued Customers: Capturing Total Customer Engagement Value', *Journal of Service Research*, 13(3), pp. 297–310.

[20] Murtagh, F. and Contreras, P. (2014) 'Algorithms for hierarchical clustering: An overview', *WIREs Data Mining and Knowledge Discovery*, 2(1), pp. 86–97.

DATA AVAILABILITY

Click Here for Repository