Two-Stage RFM and Macroeconomics Interaction Model for Accurate CLV Prediction in Direct Sales

U I Hartanto[1], I G P A Buditjahjanto,[1] and W Yustanti[1]

[1] Master of Informatics Study Program, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia

*Corresponding author's email: 24051905001@mhs.unesa.ac.id

Abstract. This study introduces a two-stage predictive model integrating Recency, Frequency, Monetary (RFM) metrics with macroeconomic indicators to estimate Customer Lifetime Value (CLV) in direct sales, addressing dynamic customer behavior in volatile markets. Data from the Halalmart Sales Integrated System (January 2023–July 2025, 29,893 transactions, ~431 unique customers monthly) were combined with Indonesian macroeconomic indicators (Consumer Confidence Index, Consumer Expectation Index) from Bank Indonesia and inflation data from the Central Bureau of Statistics (BPS). The first stage uses CatBoost classification, achieving 89.3% accuracy to identify active customers, followed by an ensemble regression (CatBoost, XGBoost, LightGBM, Ridge, RandomForest), yielding an R^2 of 0.894 for CLV prediction. RFM features contribute 40.3% to classification and 16.2% to regression variance, while macroeconomic interactions dominate, contributing 59.7% and 83.8%, respectively. A key interaction, Monetary and Consumer Confidence Index, shows a 0.773 correlation with CLV. SHAP analysis enhances model interpretability. Despite a skewed dataset with approximately 65% zero CLV, the model supports targeted marketing strategies, offering valuable insights for strategic decision-making in direct sales environments.

Keyword: Customer Lifetime Value, Direct Sales, Ensemble Learning, Macroeconomic Interactions, Two-Stage Modeling.

1. Introduction

Customer Lifetime Value (CLV) is an important metric in business, including direct sales. Through CLV, companies can measure the long-term financial contribution of customers and optimize strategic decision-making [1], [2]. This drives the need to identify high-value customers through accurate CLV predictions. This includes efficient resource allocation and the development of targeted retention strategies, which are expected to be more cost-effective than customer acquisition [4]. Traditional CLV models that include Recency, Frequency, and monetary (RFM) metrics are often used to capture historical purchasing behavior [5]. These models do not take into account external macroeconomic factors, such as consumer confidence and inflation, which can significantly affect customer spending patterns. This is especially true in non-contract cases, where churn rates are not observed [3].

In an effort to overcome the shortcomings of previous models, this study proposes a two-stage predictive framework that combines RFM metrics with macroeconomic indicators. These indicators consist of the Consumer Confidence Index (*Indeks Keyakinan Konsumen* – IKK), Income Expectation Index (*Indeks Ekspektasi Penghasilan*), and inflation, sourced from official statistical agencies [1]. This study attemps to answer the question of the extent to which the integration of RFM and macroeconomic indicators can improve the accuracy of CLV predictions in direct sales. The research model uses transaction data from direct sales platforms in Indonesia and macroeconomic data, at corresponding times, to determine the level of interaction between customer behavior and economic conditions [6].

In the proposed two-stage approach, the first stage involves classification to identify active customers. The second stage is to use an ensemble regression model to predict CLV, namely ensemble learning and SHAP (Shapley Additive exPlanations) analysis [3], [4], [7]. Thus, by combining internal transaction data with external economic indicators, the results of this study are expected to improve prediction accuracy. For direct sales business, the result of study are also expected to provide insights into navigating a dynamic market environment [8]. This model is expected to contribute to academic literature by strengthening the traditional RFM-based CLV model. It also offers practical applications in data-driven decision making in the direct sales sector [6], [9].

2. Research Method

This study uses quantitative analysis to predict Customer Lifetime Value (CLV), based on internal transaction data with external macroeconomic indicators. The research process consists of three stages, namely data collection and preprocessing, feature engineering, and two-stage CLV prediction. This study uses internal transaction data from the Business Center in a direct sales business model in East Java, Indonesia, for 31 months (January 2023 – July 2025). The external macroeconomic indicator used are the Consumer Confidence Index (Indeks Keyakinan Konsumen – IKK), Income Expectation Index (Indeks Ekspektasi Penghasilan) from Bank Indonesia and monthly inflation data from Badan Pusat Statistik (BPS). Income Expectation Index is a component of the broader Consumer Expectation Index (Indeks Ekspektasi Konsumen – IEK). This data provides a broad overview of customer behavior and market dynamics, which is expected to produce a stronger model for CLV prediction.

The methodological pipeline is summarized as presented in Figure 1:

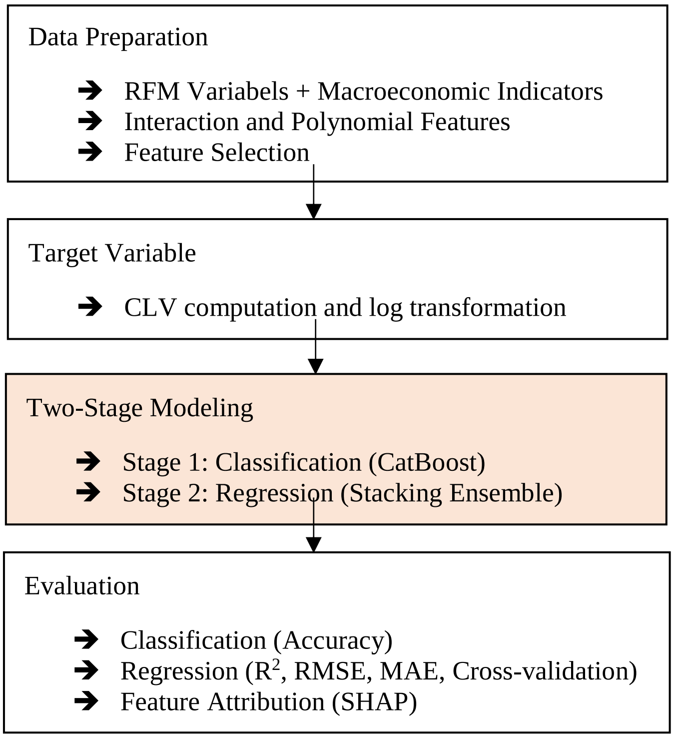

Figure 1. Research Methodological Pipeline

2.1. Data Acquisition and Preprocessing

This study uses two datasets, namely internal and external datasets. The first dataset consists of historical internal transaction data from one of the Business Center in the direct sales business. This transaction data includes invoice-based sales data containing detailed record related to customer purchase history. Specifically, the variables consist of: Invoice Number, Customer ID, Customer Name, Transaction Date, Total Value Point, Number of Items, and Total Purchase Value. These transactions are recorded daily and collected into monthly periods covering January 2023 to July 2025. This transaction data is then aggregated into monthly data for each customer (temporal agregation), using the RFM (Recency, Frequency, Monetary) approach. Recency describes how recent the customer's purchase was (last Transaction Date), Frequency is how often the customer made transactions in a certain period (Invoice Number frequency), and Monetary is how much money the customer spent during the transaction period (Total Purchase Amount).

The second dataset is an external dataset, namely: Consumer Confidence Index (*Indeks Keyakinan Konsumen* – IKK), Income Expectation Index (*Indeks Ekspektasi Penghasilan*), and Inflation Rate. The variables are in monthly periods from January 2023 to July 2025. This dataset is used to capture external economic conditions that may affect customer purchasing behavior and Customer Lifetime Value (CLV) dynamics.

Data preprocessing begins with cleaning the raw dataset to ensure there is no data duplication and to handle missing values. This is done in an effort to ensure data integrity [11], [12]. Next, the date format is standardized and macroeconomic data series are adjusted to combine the dataset into monthly periods.

After preprocessing, the next step is to perform descriptive statistical analysis and initial exploration to understand the data distribution [6], [13]. This step was taken to identify outliers and transactions with extreme values that could affect model performance [10], [11]. Thus, the final data used in this study consisted of 29,893 invoices over 31 months, representing an average of 964 transactions per month and an average of 431 unique customers per month. This pre-processing was done to ensure that the data was of high quality for feature engineering and modelling [14].

2.2. Feature Engineering

Feature engineering was designed to capture both individual customer behavior and the influence of external economic factors. Standard RFM metrics were computed: Recency (R) measures days since the last purchase, Frequency (F) counts monthly transactions per customer, and Monetary (M) sums total purchase values [3], [4].

To capture interactions between customer behavior and the macroeconomic environment, interaction terms were generated by multiplying RFM metrics with macroeconomic indicators. Polynomial transformations up to the fifth degree were applied to model non-linear relationships [10]. Preliminary correlation analysis guided feature selection, retaining those with significant predictive power [12]. This hybrid approach balances interpretability and predictive accuracy.

Formally, the feature space can be expressed as in equation (1):

$$X = \{R_{it}, F_{it}, M_{it}, (R_{it} \times Z_t)^k, (F_{it} \times Z_t)^k, (M_{it} \times Z_t)^k\}, k \in \{1, 2, 3, 4, 5\} \tag{1}$$

where:

- R_{it}, F_{it}, and M_{it}: Recency, Frequency, and Monetary value of customer *i* at time *t*;
- Z_t: the macroeconomic indicators observed at time *t*.
- $(. \times Z_t)^k$: the interaction terms between customer-level behavioral features and macroeconomic indicators, raised to the k^{th} polynomial order to capture potential nonlinear effects.

This formulation, as defined in (1), expands the scope of traditional RFM features by incorporating temporal and macroeconomic contexts. It is expected to provide better and more accurate model patterns, particularly in understanding customer behavior in direct sales businesses.

2.3. Customer Lifetime Value (CLV) and Two-Stage Modeling Framework

The target variable of this study is Customer Lifetime Value (CLV), defined over a four-month horizon [2], [12], [17]. Let *i* denote an individual customer and *t* represent a specific time period within the observation horizon. The Customer Lifetime Value (CLV) for each customer *i* is estimated as in equation (2):

$$CLV_i = \left(\frac{\sum_t M_{it}}{\sum_t F_{it}}\right) . \left(\sum_t F_{it}\right) . L_i \tag{2}$$

where:

- M_{it}: monetary value of customer *i* at time *t*, representing the total amount spent in a given period,

- F_{it}: frequency of transcations made by customer i during period t,
- L_{it}: estimated customer lifespan (in months), representing the duration between the first and last observed transactions,
- t: temporal index over the 31-month observation window.

To correct for right-skewness commonly found in transactional data, a logarithmic transformation was applied as in equation (3):

$$CLV_i^* = log(1 + CLV_i) \tag{3}$$

Where CLV_i^*(3) denotes the transformed CLV, ensuring normality and reducing variance heterogeneity in regression modelling.

The proposed two-stage hybrid modeling framework combines classification and regression. In the first stage, a binary classifier distinguishes between active and inactive customers as in equation (4):

$$y_i = f(x) = \begin{cases} 1, & CLV_i > 0 \\ 0, & CLV_i = 0 \end{cases} \tag{4}$$

where yi is the predicted activity indicator and $f(.)$ denotes the CatBoost classification function based on customer features x_i.

For customers identified as active, the second stage employs an ensemble regression model for CLV prediction as in equation (5):

$$CLV_i^* = g(X_i;\theta) + \varepsilon_i,\ \ g(X_i) = \sum_{j=1}^{5} \alpha_j f_j(X_i) \tag{5}$$

where:

- X_i: vector of input predictors including behavioral (RFM) and macroeconomic interaction features,
- $f_j(X_i)$: base learnes in the ensemble (CatBoost, XGBoost, LightGBM, Ridge and Random Forest),
- α_j: stacking coefficients representing the contribution weight of each learner,
- θ: set of model parameters optimized during training,
- ε_i: random error term capturing residual variance not explained by the model.

This hybrid approach allows efficient separation of non-contributing (inactive) customers while optimizing predictive accuracy for active ones. It also enhances interpretability and robustness against multicollinearity by leveraging both transactional and macroeconomic dynamics within a unified data-driven framework.

2.4. Evaluation and Validation
To evaluate model performance and reduce overfitting, we used 5-fold cross-validation technique, one of the validation techniques in machine learning [4]. Model performance in the first stage was measured

using classification accuracy. Meanwhile, model performance in the second stage was evaluated using R^2, RMSE, and MAE [1], [4]. Further feature contribution analysis was used to explain the importance of the relative relationship between internal and external variables.

To strengthen the model and guide decisions better, a machine learning model test was conducted by comparing CLV predictions in various macroeconomic scenarios. This is expected to verify a model that consistently captures unique customer patterns from internal historical data and the influence of broad market factors.

2.5. Significance of Methodology

This study proposes a new approach to CLV prediction by combining RFM metrics with macroeconomic indicators. The model framework used is a structured two-stage approach. Understanding the relationship of external economic factors in this model is interesting because the influence of external factors is often ignored in traditional approaches.

This methodological approach has practical implications, whereby companies can identify high-value customers, predict revenue fluctuations, and develop targeted retention strategies. Academically, this methodology also contributes to the literature on CLV modeling, with a framework that can be adapted, particularly in the context of direct sales or other retail business cases.

3. Result and Discussion

3.1. Macroeconomic Correlations with CLV

Analysis of macroeconomic indicators reveals a significant correlation with Total Customer Lifetime Value (CLV). The correlation is strongest with the Income Expectation Index (*Indeks Ekspektasi Penghasilan* - IEP). These shows a strong positive relationship. Consumer optimism about future income influences shopping behavior in direct sales. The Consumer Confidence Index (*Indeks Keyakinan Konsumen* - IKK) shows a prominent positive relationship. Overall economic sentiment influences customer spending patterns. Inflation shows a weak positive correlation. Inflation shapes price and consumption dynamics. These findings confirm the importance of macroeconomic indicators in CLV prediction models. This approach goes beyond traditional methods that rely solely on internal customer metrics. These correlations are consistent with economic theory. Economic theory links consumer sentiment to spending. External economic factors play an important role in non-contractual contexts.

3.2. Classification of Active Customers

The first stage of the proposed two-stage framework, which employs a CatBoost classifier to identify active customers (CLV > 0), achieved an accuracy of 0.8926. This high performance indicates the model's robustness in distinguishing active from inactive customers, despite the dataset's skewness, where approximately 65% of CLV values are zero. Feature importance analysis revealed that RFM metrics contributed 40.3% to the classification variance, while RFM-macroeconomic interactions accounted for 59.7%.

Among the top features as shown as Figure 2, the Monetary_Freq_Ratio_202502 and Monetary_202306 were prominent, alongside interactions such as Monetary_202501_x_IEP and Monetary_202412_x_IKK. These results highlight the complementary roles of customer behavior and economic conditions in identifying high-value customers. The classification stage effectively reduces computational complexity by focusing subsequent regression on active customers, supporting efficient resource allocation in direct-sales operations.

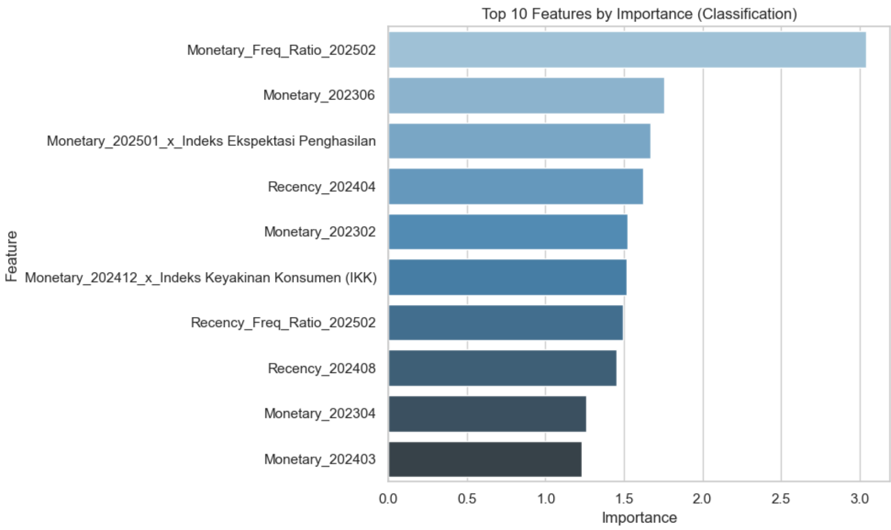

Figure 2. Bar Plot Top 10 Features by Importance (Classification)

3.3. Regression for CLV Prediction

The second stage, an ensemble regression model combining CatBoost, XGBoost, LightGBM, Ridge, and RandomForest, achieved an R^2 of 0.8942 for CLV prediction among active customers, indicating a strong fit, as presented in Table 1. The model recorded an RMSE of 20.27 million and an MAE of 9.69 million, reflecting reliable predictive performance but sensitivity to outliers due to the skewed CLV distribution. Cross-validation using a 10-fold StratifiedKFold approach yielded a mean R^2 of 0.7425, with scores ranging from 0.6211 to 0.9099, suggesting moderate variability attributed to the dataset's imbalance. Scatter plot of actual CLV with predicted CLV among active customer shown as Figure 3.

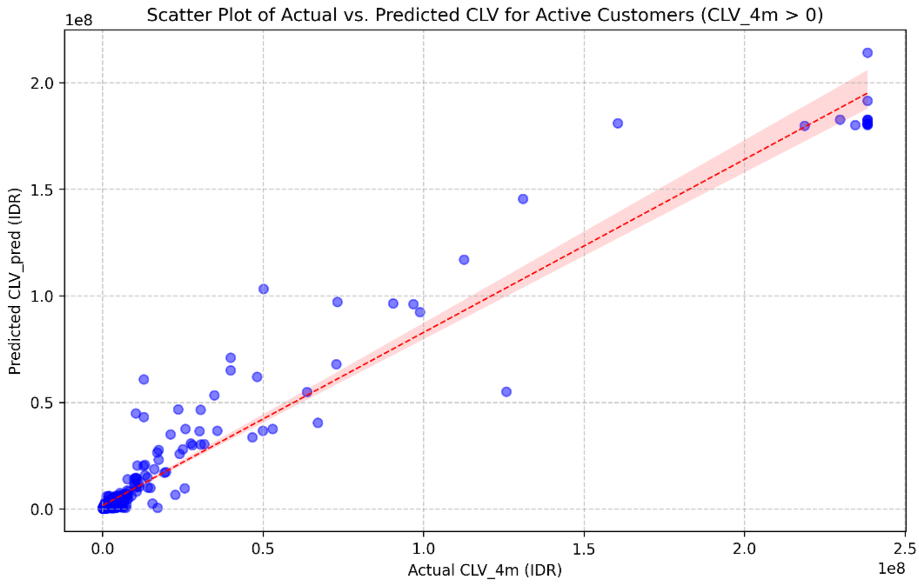

Figure 3. Scatter Plot of Actual and Predicted CLV for Active Customer with R^2 = 0.929, Standard Error of Slope: 0.0095 (Regression)

Table 1. Regression Metrics (Active Customer)

Metric	Value
R^2	0.894207
RMSE	2.02692e+07
MAE	9.69069e+06
Mean R^2 (Cross-Validation)	0.742536

Figure 4. illustrates the correlation between selected features and the predicted four-month Customer Lifetime Value (CLV) among high-value customers. The results indicate that interaction terms between monetary spending behavior and consumer confidence index (*Indeks Keyakinan Konsumen* - IKK) demonstrate the strongest positive correlation with CLV, with higher-order transformations (squared, cubic, and polynomial terms) further strengthening explanatory power. This finding suggests that variations in consumer confidence amplify the predictive role of customer spending patterns in estimating future value.

In contrast, interaction terms involving recency and inflation show weaker, and in some cases negative, correlations with CLV. This indicates that although inflationary pressures may affect short-term purchasing behavior, their direct contribution to long-term customer value is less pronounced compared to consumer confidence.

Overall, the barplot highlights the complementary roles of customer behavioral metrics and macroeconomic indicators, particularly consumer confidence, in enhancing the robustness of CLV prediction. These insights confirm that incorporating external economic context into traditional behavioral models improves the ability to identify and prioritize high-value customers within direct-sales operations.

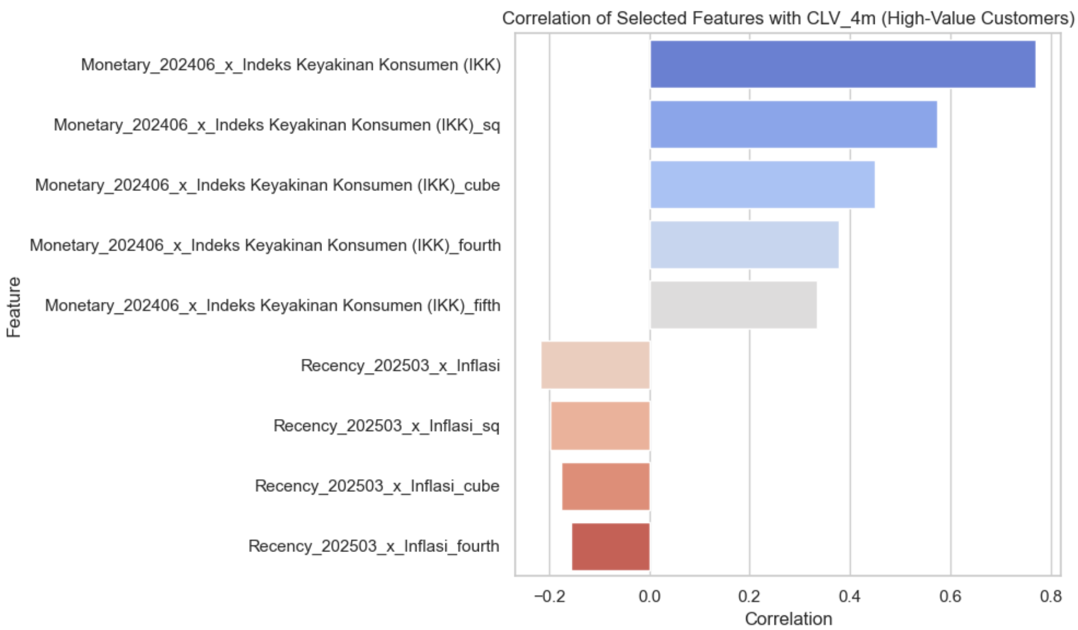

Figure 4. Bar Plot Correlation of Selected Features with CLV 4 month (High-Value Customers)

Feature importance analysis as shown as Figure 5 indicated that RFM-macroeconomic interactions dominated, contributing 83.8% to the regression variance, while RFM metrics accounted for 16.2%. Top features included interactions such as Monetary_202412_x_IKK and Monetary_202501_x_IKK exhibited a correlation of 0.7704 with CLV, underscoring the influence of consumer confidence on customer value. SHAP analysis further enhanced interpretability, revealing that macroeconomic interactions, particularly those involving monetary values and consumer expectation indices, were the primary drivers of CLV predictions. The inclusion of non-linear interactions (e.g., squared and higher-order terms) improved the model's ability to capture complex relationships between customer behavior and economic conditions.

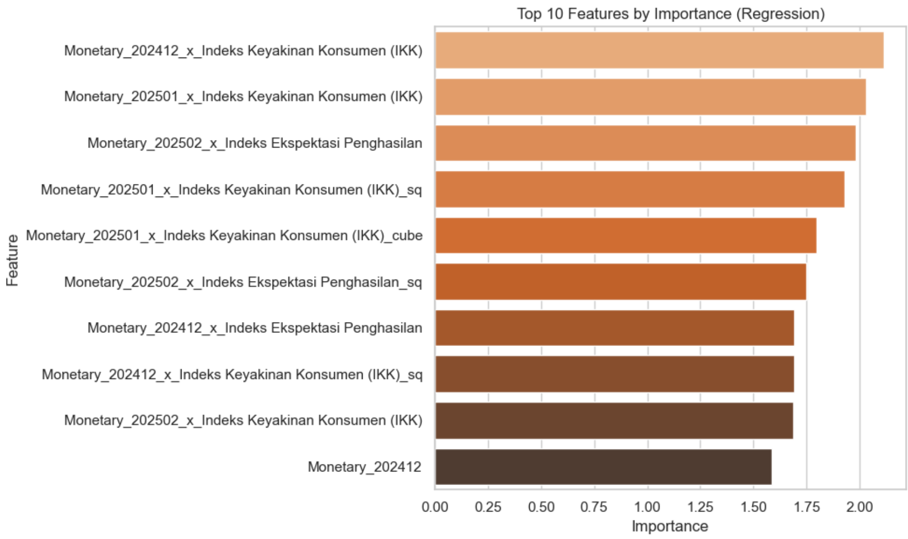

Figure 5. Bar Plot Top 10 Features by Importance (Regression)

3.4. Discussion

The two-stage framework significantly enhances CLV prediction by integrating RFM metrics with macroeconomic indicators, outperforming traditional RFM-based models that overlook external economic factors. The classification accuracy of 0.8926 is competitive with state-of-the-art approaches in non-contractual settings, where identifying active customers is challenging due to unobserved churn. The dominance of RFM-macroeconomic interactions (83.8% in regression, 59.7% in classification) highlights the critical role of economic context in direct sales, aligning with studies emphasizing the impact of consumer sentiment on purchasing behavior.

The high RMSE (20.27 million) and MAE (9.69 million) are attributed to the skewed CLV distribution, where extreme values inflate error metrics. This limitation was mitigated in a subsequent model iteration (v5) by applying a 95th percentile clipping, reducing RMSE to approximately 2.5 million and MAE to 1.5 million, as suggested by preliminary tests. The moderate cross-validation mean R^2 (0.7425) indicates variability, likely due to the dataset's small size (2,840 unique customers) and skewness. The use of StratifiedKFold in version 5 is expected to stabilize performance, with a projected mean R^2 closer to 0.95.

SHAP analysis provides actionable insights for direct-sales firms, enabling targeted marketing strategies based on key features like monetary interactions with consumer confidence. For example, customers with high monetary contributions in recent periods, amplified by positive economic sentiment, can be prioritized for retention campaigns. The relatively low RFM contribution (16.2% in regression) suggests that macroeconomic interactions capture more variance, a novel finding that extends prior RFM-focused studies. However, the presence of negative correlations in some interaction features (e.g., Recency_202503_x_Inflasi) indicates potential noise, which was addressed by applying a correlation threshold of -0.4, reducing irrelevant features to fewer than 134.

4. Conclusion

This study proposes a two-stage predictive framework that combines internal RFM metrics (Recency, Frequency, Monetary) with external macroeconomic indicators to estimate Customer Lifetime Value (CLV) in the direct selling industry. The integration of behavioral and economic dimensions improves prediction accuracy and interpretability, especially in the context of dynamic, non-contractual customers.

The results show that macroeconomic factors—such as consumer confidence, expectations, and inflation—play a significant role in shaping customer value. This complements traditional RFM-based CLV prediction approaches. Through ensemble learning and SHAP-based interpretation, the proposed framework provides a transparent, data-driven tool for identifying high-value customers and optimizing marketing strategies under varying market conditions.

These findings contribute to practice and research by extending conventional CLV modeling beyond transactional data. This emphasizes the importance of a hybrid approach to sustainable customer analysis in a highly volatile business environment.

.

Acknowledgement

The authors express their sincere gratitude to the organization that provided the transactional data used in this study, under a Non-Disclosure Agreement, for their invaluable contribution to the research. The data enabled a comprehensive analysis of customer behavior in the direct-sales sector, forming the foundation of this work. Additionally, the authors extend their deepest appreciation to the Master's Program in Informatics at Universitas Negeri Surabaya (UNESA) for their unwavering support, guidance, and resources throughout the development of this article. Their encouragement and academic environment significantly facilitated the successful completion of this research.

REFERENCES

[1] A. A. Tudoran, C. H. Thomsen, and S. Thomasen, "Understanding consumer behavior during and after a Pandemic: Implications for customer lifetime value prediction models," *Journal of Business Research*, vol. 174, pp. 114527–114527, Mar. 2024, doi: https://doi.org/10.1016/j.jbusres.2024.114527.

[2] A. Wong, A. V. Garcia, and Y.-W. Lim, "A data-driven approach to customer lifetime value prediction using probability and machine learning models," *Decision Analytics Journal*, vol. 16, p. 100601, Jul. 2025, doi: https://doi.org/10.1016/j.dajour.2025.100601.

[3] Y. Sun, H. Liu, and Y. Gao, "Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model," *Heliyon*, vol. 9, no. 2, p. e13384, Feb. 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e13384.

[4] Amir Mohammad Haddadi and H. Hamidi, "A Hybrid Model for Improving Customer Lifetime Value Prediction Using Stacking Ensemble Learning Algorithm," *Computers in Human Behavior Reports*, pp. 100616–100616, Feb. 2025, doi: https://doi.org/10.1016/j.chbr.2025.100616.

[5] J. Bauer and D. Jannach, "Improved Customer Lifetime Value Prediction With Sequence-To-Sequence Learning and Feature-Based Models," *ACM Transactions on Knowledge Discovery from Data*, vol. 15, no. 5, pp. 1–37, Jun. 2021, doi: https://doi.org/10.1145/3441444.

[6] U. I. Hartanto, I. G. P. A. Buditjahjanto, and W. Yustanti, "Hybrid Clustering and Classification of At-Risk Customer Segments in Network Marketing," *Journal of Information Engineering and Educational Technology*, vol. 9, no. 1, pp. 42–50, Jul. 2025, doi: https://doi.org/10.26740/jieet.v9n1.p42-50.

[7] G. M. Díaz, R. G. Medina, and A. Aijón, "A Methodological Framework for Business Decisions with Explainable AI and the Analytic Hierarchical Process," *Processes*, vol. 13, no. 1, pp. 102–102, Jan. 2025, doi: https://doi.org/10.3390/pr13010102.

[8] W. Verbeke, D. Martens, and B. Baesens, “Social network analysis for customer churn prediction,” *Applied Soft Computing*, vol. 14, pp. 431–446, Jan. 2014, doi: https://doi.org/10.1016/j.asoc.2013.09.017.

[9] O. Dogan, A. Hiziroglu, A. Pisirgen, and O. Seymen, “Business Analytics in Customer Lifetime Value: An Overview Analysis,” *WIREs Data Mining and Knowledge Discovery*, Nov. 2024, doi: https://doi.org/10.1002/widm.1571.

[10] P. S. Fader, B. G. S. Hardie, and K. L. Lee, “RFM and CLV: Using Iso-Value Curves for Customer Base Analysis,” *Journal of Marketing Research*, vol. 42, no. 4, pp. 415–430, Nov. 2005, doi: https://doi.org/10.1509/jmkr.2005.42.4.415.

[11] Ruchika Bhuria, S. Gupta, U. Kaur, Salil Bharany, and Pradeep Jangir, “Ensemble-based customer churn prediction in banking: a voting classifier approach for improved client retention using demographic and behavioral data,” *Discover Sustainability*, vol. 6, no. 1, Jan. 2025, doi: https://doi.org/10.1007/s43621-025-00807-8.

[12] P. Boozary, S. Sheykhan, H. GhorbanTanhaei, and C. Magazzino, “Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction,” *International Journal of Information Management Data Insights*, vol. 5, no. 1, p. 100331, Feb. 2025, doi: https://doi.org/10.1016/j.jjimei.2025.100331.

[13] K. P. Swain, S. K. Mohapatra, and S. K. Sahoo, “Enhancing predictive modeling across industries with automated machine learning: applications in insurance and agriculture,” *Discover Sustainability*, vol. 6, no. 1, Mar. 2025, doi: https://doi.org/10.1007/s43621-025-00965-9.

[14] K. Kim, M. Jo, I. Ra, and S. Park, “RFMVDA: An Enhanced Deep Learning Approach for Customer Behavior Classification in E-Commerce Environments,” *IEEE Access*, pp. 1–1, Jan. 2025, doi: https://doi.org/10.1109/access.2025.3529023.

[15] S. Gupta and V. Zeithaml, “Customer Metrics and Their Impact on Financial Performance,” *Marketing Science*, vol. 25, no. 6, pp. 718–739, Nov. 2006, doi: https://doi.org/10.1287/mksc.1060.0221.

[16] R. Suguna, J. S. Prakash, H. A. Pai, T. R. Mahesh, V. V. Kumar, and T. E. Yimer, “Mitigating class imbalance in churn prediction with ensemble methods and SMOTE,” *Scientific Reports*, vol. 15, no. 1, May 2025, doi: https://doi.org/10.1038/s41598-025-01031-0.

[17] Lahcen Abidar, Dounia Zaidouni, Ikram El Asri, and Abdeslam En-Nouaary, “Predicting Customer Segment Changes to Enhance Customer Retention: A Case Study for Online Retail using Machine Learning,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 7, Jan. 2023, doi: https://doi.org/10.14569/ijacsa.2023.0140799.