Perbanas International Conference on Economics, Business, Management, Accounting and IT (PROFICIENT) 2026
"Integrity-Driven Innovation: Reimagining Business, Finance, and Digital Transformation for Sustainable Impact"

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Customer Lifetime Value Prediction Using Linear Regression to Support CRM Strategy in Banking Industry

Sri Wahyuni Aprianti[1,a], Yoga Armando[2,b], Alrafiful Rahman[3,c]

[1,2,3] Data Science Study Program, Faculty of Information Technology, Perbanas Institute

[a)] Corresponding author: sri.wahyuni06@perbanas.id,
[b)]yoga.armando08@perbanas.id,
[c)]alrafiful.rahman@perbanas.id

Abstract *- This study aims to predict Customer Lifetime Value (CLTV) using a Linear Regression approach to support Customer Relationship Management (CRM) strategies in the banking industry. The research utilizes customer transaction and demographic data to identify key factors influencing customer value. The workflow includes data preparation, exploratory data analysis, preprocessing, model development, and evaluation. Variables such as age, transaction frequency, balance, and product usage are analyzed as predictors of CLTV. The Linear Regression model is selected due to its interpretability and ability to explain relationships between variables. Model performance is evaluated using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results show that several variables significantly affect customer value and can be used to estimate future contributions of customers. The predicted CLTV is then utilized to support CRM strategies, including customer segmentation, personalized product offerings, and retention programs. This study demonstrates that a data-driven approach can enhance decision making and improve service effectiveness in banking. The findings provide practical insights for financial institutions in optimizing customer management and increasing long term profitability. Furthermore, the model supports strategic planning by enabling early identification of high value customers and improving allocation of marketing resources effectively and efficiency*

Keywords: Customer Lifetime Value (CLTV), Linear Regression, Customer Relationship Management (CRM), Banking Industry, Customer Segmentation, Data-Driven Decision Making

INTRODUCTION

In the era of digital transformation, the banking industry is increasingly adopting a customer-centric approach to maintain competitiveness and profitability. Customer Lifetime Value (CLTV) is an important metric used to measure the total profits generated by customers and support effective marketing and retention strategies (Oghenemaro, 2025). By utilizing CLTV, banks can optimize resources, increase customer loyalty, and improve overall financial performance (Egorenkov, 2024). Traditional methods for evaluating customer value often can't accurately capture complex customer behavior. (Oghenemaro, 2025). Linear regression is widely used for CLTV predictions due to its simplicity and interpretation.

The integration of CLTV predictions in CRM strategies helps banks identify high-value customers and increase customer loyalty. CLTV is considered an important customer-centric metric that supports customer retention and satisfaction through the implementation of a data-driven CRM (Kuštelega et al., 2025). Predictive CRM analytics help organizations turn customer data into actionable insights to improve retention and marketing strategies. Combining predictive modeling with a CRM system allows for a more effective and data-driven business environment (Mist, 2024). The proposed model is expected to improve customer segmentation, decision-making, and long-term profitability.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

LITERATURE REVIEW

Customer Lifetime Value (CLTV)

Customer Lifetime Value (CLTV) is an important concept in measuring customer long-term economic value and supporting Customer Relationship Management (CRM) strategies in the banking industry through service personalization and predictive decision-making ((Oghenemaro, 2025);(Egorenkov, 2024)). However, the literature review related to CLTV still tends to be descriptive and has not provided a critical synthesis of the development of modern predictive methods. Therefore, this study uses Linear Regression because it is able to provide a clear interpretation of the relationships between variables so that it can support strategic decision-making in banking CRM (Sun et al., 2023).

Linear Regression in Customer Value Prediction

Multiple linear regression is a predictive method that is widely used in the business and financial sectors because it is simple, easy to interpret, and able to explain the relationship between independent variables and dependent variables (Gadgil et al., 2024). This model is formulated as

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon.$$

and used to predict CLTV based on variables such as transaction frequency, average balance, customer age, and number of products by testing classical assumptions to ensure model validity (Montgomery et al., 2021).

Customer Relationship Management (CRM) in the Banking Sector

Customer Relationship Management (CRM) is a strategy that integrates business processes and technology to build long-term relationships with customers and increase customer loyalty in the banking sector (Kumar et al., n.d., 2018). CRM leverages transaction and customer behavior data to support service personalization, customer segmentation, retention strategies, and CLTV integration to improve the effectiveness of data-driven decision-making and the accuracy of high-value customer predictions (Kuštelega et al., 2025).

Value-Based Customer Segmentation in Digital Banking

Data-based customer segmentation is used to group customers based on characteristics and estimated future mass value so that marketing and service strategies become more targeted (Lavanya & Bharathi, 2022). In banking, CLTV-based segmentation helps banks identify high-value customers, hence the risk of churn, and supports service personalization through integration with CRM systems (Kumar et al., n.d., 2018).

Previous Research

Previous studies have shown that machine learning techniques can be effectively applied in the banking industry to support predictive analytics and Customer Relationship Management (CRM). One relevant implementation was conducted using the "ML DDE CIMB Niaga Demo" dataset published on Kaggle by Dyah Nurlita. The implementation demonstrated the use of machine learning methods for customer data analysis, including data preprocessing, exploratory data analysis, regression modeling, and model evaluation using MAE and MAPE metrics.

The study showed that transactional customer variables had significant influence on prediction results and could support data-driven decision making in the banking sector. In addition, predictive analytics helped improve customer segmentation, retention strategies, and personalized service recommendations. Therefore, this research extends previous work by focusing specifically on Customer Lifetime Value (CLTV) prediction using Linear Regression to support CRM strategies in the banking industry.

Research Dataset: CIMB Niaga

This study uses CIMB Niaga's Kaggle Demo ML DDE public dataset containing transactional and demographic data of Indonesian banking customers to support Customer Lifetime Value (CLTV) modeling Dyah Nurlita. The dataset includes variables such as age, frequency and value of transactions, product usage, and digital

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

service engagement so as to provide an overview of customer behavior that is relevant to the Indonesian banking context.

Research Conceptual Framework

The following conceptual framework explains the process of Customer Lifetime Value (CLTV) prediction using Linear Regression to support CRM strategy in the banking industry.

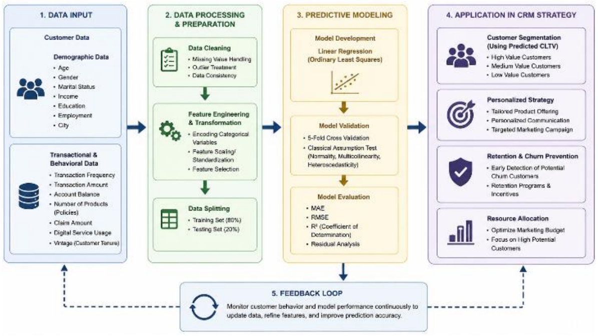

FIGURE 1. Conceptual Framework of Customer Lifetime Value Prediction Using Linear Regression

As shown in Figure 1, the research framework begins with customer demographic and transactional data collection, followed by data preprocessing and feature engineering processes to improve data quality and model readiness. The processed data is then used to develop and evaluate the Linear Regression model for predicting Customer Lifetime Value (CLTV). The prediction results are integrated into CRM strategies such as customer segmentation, personalized marketing, retention programs, and resource allocation. In addition, the framework includes a feedback loop mechanism to continuously improve prediction accuracy and customer relationship management effectiveness.

METHODS

Design and Research Flow

This study uses a quantitative approach with a predictive study design to build a linear regression model in predicting Customer Lifetime Value (CLTV) based on customer transaction characteristics and behavior. The research process includes data collection, exploratory analysis, pre-processing and feature engineering, model development, validation, and interpretation of results to support CRM strategies.

Data Sources

This study uses the Kaggle public dataset "ML DDE CIMB Niaga Demo" which contains transactional and demographic data of CIMB Niaga Indonesia Bank customers. The dataset consists of 73,637 rows and 10 columns, making it suitable for Customer Lifetime Value (CLTV) prediction analysis. Several variables used in this study include customer age, number of policies, claim amount, vintage, transaction-related variables, and other customer behavioral attributes that support the development of the Linear Regression model. The dataset was selected because it is relevant to the Indonesian banking context and provides sufficient information to support predictive analytics and CRM strategy implementation.

Pre-Process and Feature Engineering

The research data goes through the pre-processing stage to ensure quality and readiness before modeling is carried out. This process includes handling missing values, identifying outliers, normalization, standardization, and feature engineering to increase the relevance of variables in CLTV prediction, where outliers are retained because they can represent high-value customers.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is performed to understand the structure, distribution, and relationships between variables before the model is built. This process uses descriptive statistics and visualizations such as histograms, scatter plots, boxplots, and correlation heatmaps to support feature selection and detect potential multicollinearity.

Data Splitting

The dataset was divided into training data and testing data using the train_test_split() function from the Scikit-learn library with an 80:20 ratio. Approximately 80% of the dataset was used for training the Linear Regression model, while the remaining 20% was used to evaluate the model's predictive performance on unseen data. The parameter random_state = 42 was applied to ensure experiment reproducibility and maintain consistent results during the model training and evaluation process.

Development of Linear Regression Models

The quantitative workflow using the Ordinary Least Squares (OLS) method has been systematically compiled through assumption validation, variable selection, and cross-validation. However, the use of basic linear models on Customer Lifetime Value (CLTV) data that have skewed distributions and complex patterns of customer behavior can be limited in capturing non-linear relationships between variables. Several studies have shown that modern machine learning methods are capable of providing higher predictive performance, although they often come at the expense of model interpretability ((Richman & Wüthrich, 2023); (Spyrison et al., 2025)).

Model Evaluation

Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and determination coefficient (R^2) to measure prediction accuracy and model's ability to explain CLTV variations. Residual analysis and interpretation of regression coefficients were also performed to identify error patterns and variables that had the most influence on CLTV values to support a more effective CRM strategy.

Implementation in CRM Strategy

CLTV's predictive results are integrated into the CRM strategy through customer segmentation based on customer value to support retention strategies and more targeted services. In addition, the results of the model are used for personalization of product offerings and early detection of potential churn so that banks can intervene proactively (Judijanto et al., 2026).

RESULTS AND DISCUSSION

System Implementation Overview

The Customer Lifetime Value (CLTV) prediction system was implemented using the Python programming language within the Jupyter Notebook environment. Several libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, and Statsmodels were utilized to support data preprocessing, visualization, regression modeling, and statistical analysis.

TABLE 1. Libraries Used in the Research

Library	Function
Pandas	Data Manipulation and prepocessing
NumPy	Numerical computation
Matplotlib	Data Visualization
Seaborn	Satistical Data Visualization
Scikit-learn	Mechine learning modeling and evaluation
Statsmodel	Statistical Regression Analysis

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

As shown in Table 1, several Python libraries were integrated to support the implementation of the Linear Regression model. These libraries enabled efficient data processing, visualization, and evaluation throughout the research process.

Dataset Exploration

The dataset used in this study contained customer demographic and transactional information related to Customer Lifetime Value (CLTV). Initial exploration was conducted using functions such as head(), info(), and isnull().sum() to understand the structure and quality of the dataset.

TABLE 2. Initial Dataset Exploration Process

Process	Purpose
head()	Display the first rows of the dataset
info()	Display dataset structure and data types
isnull().sum()	Check missing value

As shown in Table 2, the dataset exploration process successfully identified the structure, attribute types, and statistical characteristics of the dataset. The dataset was suitable for further preprocessing and predictive analysis because all variables were successfully recognized and organized.

Data Prepocessing

Data preprocessing was performed to improve dataset quality before the modeling stage. The preprocessing activities included checking missing values, encoding categorical variables, feature scaling, and splitting the dataset into training and testing data.

TABLE 3. Data Preprocessing Activities

Activity	Method
Missing Value Checking	isnull().sum()
Categorical Encoding	LabelEncoder()
Feature Scaling	StandardScaler()
Data Splitting	train_test_split()

As shown in Table 3, the preprocessing stage successfully transformed the raw dataset into a structured format suitable for machine learning analysis. Categorical variables were converted into numerical representations to support the implementation of the Linear Regression model.

Correlation Analysis

Correlation analysis was conducted to identify the relationships between variables in the dataset before the modeling stage. This analysis is important to understand how each feature contributes to Customer Lifetime Value (CLTV) prediction and to detect potential multicollinearity problems among independent variables.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

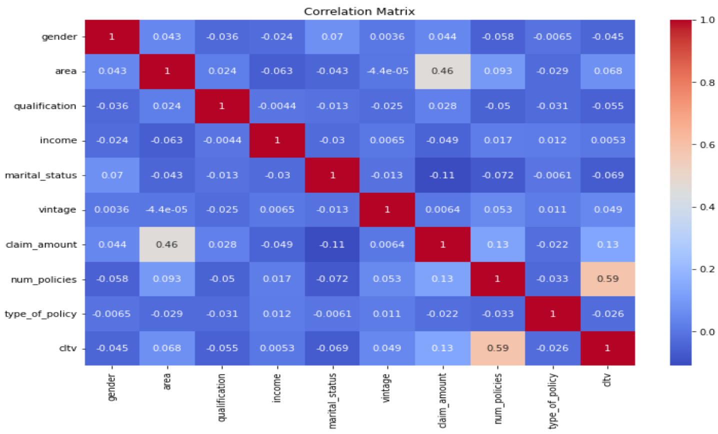

FIGURE 2. Correlation Heatmap

As shown in Figure 2, the correlation heatmap illustrates the relationships between variables in the dataset. Several transactional variables such as number of policies and claim amount showed stronger correlations with CLTV compared to demographic variables. The visualization also helped identify potential multicollinearity issues before the regression modeling stage. Overall, the correlation analysis supported feature selection and model preparation.

Model Evaluation

The Linear Regression model was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). These metrics were used to measure prediction accuracy and the model's ability to explain CLTV variations.

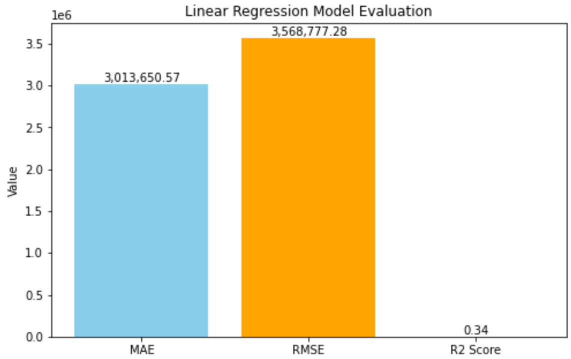

FIGURE 3. Model Evaluation Results

As shown in Figure 3, the model achieved moderate predictive performance with acceptable MAE and RMSE values. However, the relatively low R^2 value suggests that the model is still not able to optimally account for most of the variation in customer value. This indicates that the relationship pattern between predictor variables and CLTV tends to be complex and non-linear so that the Linear Regression approach has limitations in capturing the overall customer behavior. Therefore, the results of the evaluation need to be interpreted more critically considering the possibility of using more complex predictive methods to improve the model's predictive capabilities.

Feature Importance Analysis

Feature importance analysis was conducted using regression coefficient values to identify variables with the strongest influence on CLTV prediction.

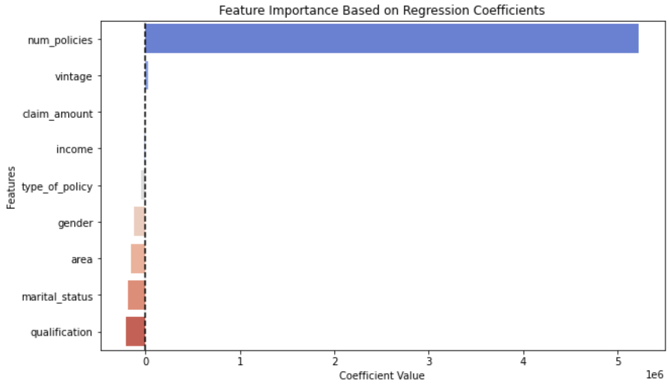

FIGURE 4. Feature Importance Based on Regression Coefficients

As shown in Figure 4, the number of policies variable had the strongest positive contribution to CLTV prediction. This finding indicates that customers with multiple banking products tend to generate higher long-term value for the company. Several demographic variables showed weaker contributions compared to transactional variables. This result confirms that customer engagement and product utilization are important indicators in predicting customer profitability.

Actual vs Predicted CLTV

Actual versus predicted analysis was conducted to evaluate the prediction capability of the Linear Regression model in estimating Customer Lifetime Value (CLTV). This visualization compares the actual customer values with the predicted results generated by the regression model.

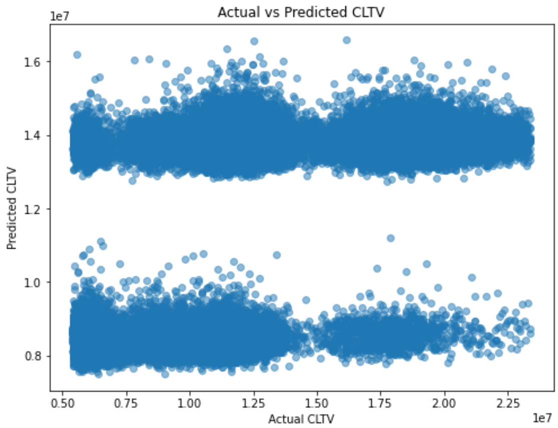

FIGURE 5. Comparison Between Actual and Predicted Customer Lifetime Value (CLTV)

As shown in Figure 5, the predicted CLTV values generally followed the distribution pattern of the actual CLTV values. Several prediction points were closely aligned with the actual observations, indicating that the model was capable of capturing overall customer value trends. However, several deviations between actual and predicted values were still identified, particularly for customers with extremely high CLTV values. This result indicates that although the Linear Regression model performed reasonably well, prediction accuracy for outlier customers remained relatively challenging.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Residual Distribution Analysis

Residual distribution analysis was conducted to evaluate the error patterns generated by the Linear Regression model. This analysis is important to determine whether the prediction errors were distributed normally and to assess the reliability of the regression model.

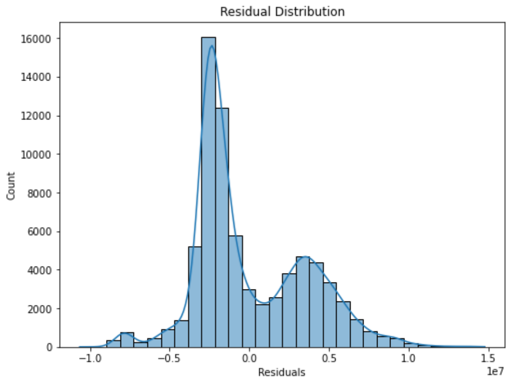

FIGURE 6. Distribution of Regression Residuals

As shown in Figure 6, the residual values were distributed relatively close to zero, indicating that the prediction errors were reasonably balanced across the dataset. The majority of residuals were concentrated around the center of the distribution, suggesting that the model produced stable prediction performance. However, slight skewness was still observed in the residual distribution, indicating that several customer values were difficult to predict accurately. Overall, the residual analysis confirms that the Linear Regression model produced acceptable error patterns for Customer Lifetime Value (CLTV) prediction.

OLS Regression Analysis

Ordinary Least Squares (OLS) regression analysis was conducted to evaluate the statistical significance of the regression model and its predictor variables.

TABLE 4. OLS Regression Summary

Activity	Method
R^2	0.3448
Adjusted R^2	0.3447
Prob (F-statistic)	< 0.05
Model Type	Linear Regression

As shown in Table 4, the regression model was statistically significant with a Prob (F-statistic) value below 0.05. This result indicates that the independent variables jointly contributed significantly to CLTV prediction. The OLS analysis also confirmed that several variables such as number of policies, vintage, and claim amount had meaningful influence on customer lifetime value. These findings support the effectiveness of Linear Regression for customer profitability analysis in the banking sector.

DISCUSSION

The results of this study demonstrate that the Linear Regression model was able to provide meaningful predictions of Customer Lifetime Value (CLTV) using customer demographic and transactional data from the banking sector. The model evaluation results showed acceptable performance based on MAE, RMSE, and R^2

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

metrics, indicating that the model could explain part of the variation in customer value. Although the R^2 value was moderate, the model successfully captured important behavioral patterns related to long-term customer profitability. However, the moderate coefficient of determination also indicates that a considerable portion of customer value variation could not be fully explained by the model, which may create operational risks in strategic implementation, particularly in financial resource allocation, high-value customer segmentation, and customer retention prioritization. These findings suggest that Linear Regression can serve as an interpretable and practical approach for CLTV prediction in banking environments, particularly when organizations require transparent and explainable analytical models, although the linear approach still has limitations in capturing complex and nonlinear customer behavior patterns.

The correlation and feature importance analyses revealed that transactional variables contributed more strongly to CLTV prediction than demographic variables. In particular, the number of policies and claim amount showed significant influence on customer value. This indicates that customers who actively use multiple banking products tend to generate higher long-term profitability for the bank. The OLS regression analysis further confirmed that several independent variables significantly affected CLTV, supported by the statistically significant Prob (F-statistic) value below 0.05. These findings are consistent with previous studies emphasizing that customer engagement, transaction activity, and product utilization are important indicators in predicting customer profitability and supporting data-driven CRM strategies.

The actual versus predicted visualization and residual distribution analysis also demonstrated that the model produced relatively stable prediction patterns. Most residual values were concentrated near zero, indicating balanced prediction errors across the dataset. However, several deviations were still observed, especially for customers with extremely high CLTV values. This limitation suggests that Linear Regression may struggle to fully capture complex nonlinear relationships and extreme customer behaviors. Nevertheless, the model remains valuable for supporting CRM implementation, including customer segmentation, personalized service strategies, retention programs, and marketing resource allocation. Overall, this study confirms that integrating predictive analytics into CRM can improve strategic decision-making and help banking institutions optimize long-term customer management, although the implementation of this model in real business environments still requires careful interpretation and consideration of more advanced predictive approaches to improve prediction accuracy and CRM effectiveness.

CONCLUSIONS

This study successfully implemented a Linear Regression model to predict Customer Lifetime Value (CLTV) in the banking industry using transactional and demographic customer data. The research process included data preprocessing, exploratory data analysis, regression modeling, evaluation, and interpretation to support Customer Relationship Management (CRM) strategies. The evaluation results showed that the model achieved acceptable predictive performance and was able to explain customer value patterns through several significant predictor variables.

The findings revealed that transactional variables, particularly the number of policies and claim amount, had the strongest influence on CLTV prediction. The regression analysis also confirmed that the model was statistically significant and capable of supporting customer profitability analysis. Although prediction limitations were observed for customers with extremely high CLTV values, the overall model performance remained stable and reliable for practical business applications.

In conclusion, the integration of CLTV prediction into CRM strategies can help banks improve customer segmentation, personalize services, optimize retention programs, and allocate marketing resources more effectively. This study demonstrates that a data-driven predictive approach can support strategic decision-making and contribute to improving long-term profitability in the banking sector.

REFERENCES

Egorenkov, D. (2024). *AI-Powered Predictive Customer Lifetime Value : Maximizing Long-Term Profits*. *12*(09), 1428–1436. https://doi.org/10.18535/ijsrm/v12i09.ec02

Gadgil, K., Gill, S. S., & Abdelmoniem, A. M. (2024). lifetime value prediction. *Journal of Economy and*

Technology, *1*(September 2023), 197–207. https://doi.org/10.1016/j.ject.2023.09.001

Judijanto, L., Koraag, R., Satria, F., & Prambodo, Y. L. (2026). *INFORMATION SYSTEMS IN THE DIGITAL ERA: BUILDING A RESILIENT TECHNOLOGY INFRASTRUCTURE*. Mind Power Publishing.

Kabut, S. A. (2024). *A PREDICTIVE CRM ANALYTICS FRAMEWORK FOR MERCHANT RETENTION : APPLYING RFM SEGMENTATION , CUSTOMER PROFILING , AND BEHAVIORAL ANALYTICS IN THE B2B PAYMENT INSTITUT TEKNOLOGI BANDUNG JULY 2024 ABSTRACT A PREDICTIVE CRM ANALYTICS FRAMEWORK FOR MERCHANT RETENTION : APPLYING RFM SEGMENTATION , CUSTOMER PROFILING , AND BEHAVIORAL ANALYTICS IN THE B2B PAYMENT GATEWAY COMPANY DOKU*. *29122418*(July).

Kumar, V., Reinartz, W., & Edition, T. (n.d.). *Customer Relationship Management*.

Kuštelega, M., Gregurec, I., & Hrustek, L. (2025). *The Impact of Customer Lifetime Value in Customer Relationship Management : A Bibliometric Analysis*. *14*(3), 2711–2722. https://doi.org/10.18421/TEM143

Lavanya, R., & Bharathi, B. (2022). *Effective Cross Synthesized Methodology for Movie Recommendation with Emotion Analysis through Ranking Score*. *13*(5), 226–233.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

Oghenemaro, S. A. (2025). *Open Access Optimizing Customer Lifetime Value (CLV) Prediction Models in Retail Banking Using Deep Learning and Behavioral Segmentation*. *07*, 123–131.

Richman, R., & Wüthrich, M. V. (2023). *LocalGLMnet : interpretable deep learning for tabular data*. *1238*. https://doi.org/10.1080/03461238.2022.2081816

Spyrison, N., Cook, D., & Biecek, P. (2025). Exploring local explanations of nonlinear models using animated linear projections. *Computational Statistics*, *40*(2), 1071–1095. https://doi.org/10.1007/s00180-023-01453-2

Sun, Y., Liu, H., & Gao, Y. (2023). Heliyon Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon*, *9*(2), e13384. https://doi.org/10.1016/j.heliyon.2023.e13384