Machine Learning-Powered Customer Lifetime Value Segmentation for Predicting Customer Value in the E-commerce Industry

Airiel Bin Rudy @ Basir1 and Zaidatol Haslinda Abdullah Sani2,\*[0002-8406-2967]

1, 2 Faculty of Computing and Informatics, Universiti Malaysia Sabah, 88400, Kota Kinabalu, Malaysia

1airiel\_rudy\_bi21@iluv.ums.edu.my

2,\*linda.sani@ums.edu.my

**Abstract.** Traditional e-commerce marketing strategies are struggling to keep pace with the need for personalized customer experiences. Customer acquisition costs often outweigh the benefits of retaining existing customers, highlighting the importance of understanding long-term customer value. However, bombarding customers with irrelevant information can backfire, leading to lost sales. Many current marketing techniques lack a focus on Customer Lifetime Value (CLV), use overly simplified segmentation methods, and do not account for how customer behavior evolves. This project seeks to overcome these challenges by developing a comprehensive customer segmentation model specifically for e-commerce: K-means clustering, DBSCAN, and GMM. This project aims to combine data analysis techniques with sophisticated machine learning algorithms to predict CLV using classification models, which are Decision Tree, Random Forest, Gradient Boosting, and SVM. Customer segments will be created based on their predicted value and behavioral patterns. The project will then measure the effectiveness of the segmentation and prediction approach, evaluating each machine learning model to know the performance. The result shows that K-Means balanced simplicity and efficacy, while DBSCAN was particularly useful for detecting and removing outliers. Gradient Boosting emerged as the best-performing model, delivering high accuracy and stability across different segmentation methods. Decision Tree and Random Forest also performed well, particularly in GMM-based segmentation, while SVM showed weaker performance with lower accuracy and higher variance. Hyperparameter tuning was applied to each model to optimize performance, and evaluations were conducted using accuracy, precision, recall, F1-score, and T-tests to ensure robustness. The result of this study could offer e-commerce businesses valuable ways to target marketing, improve customer retention, and boost long-term profitability.

**Keywords:** Segmentation, Clustering, Predict, Classification.

1. Introduction

Customer Lifetime Value (CLV) is a critical metric in the e-commerce industry that quantifies the financial worth of a customer throughout their entire relationship with a business. This metric provides companies with valuable insights into the long-term contributions of their customers, facilitating informed decision-making regarding customer acquisition, retention strategies, and resource allocation [3]. Accurately predicting CLV poses significant challenges due to the dynamic nature of the e-commerce landscape and the diverse customer behavior patterns. Many e-commerce organizations struggle to effectively understand and optimize the true value of their customer base, leading to potential revenue losses and inefficient marketing strategies [5].

This project aims to develop a CLV segmentation model powered by machine learning techniques to transform client targeting into e-commerce and predict low, medium, and high customer lifetime value. Businesses can implement personalized marketing campaigns that drive sales and foster long-term consumer relationships by accurately predicting customer value and clustering customers into actionable segments. Customer segmentation involves categorizing consumers based on shared characteristics, enabling companies to effectively tailor their marketing strategies and product offerings [7]. This targeted approach enhances customer satisfaction and improves conversion rates by aligning products with consumer preferences.

Traditional segmentation methods often rely on basic purchase history data or simplistic demographic information, which may fail to capture the complexities of consumer behavior. In contrast, machine learning-driven segmentation can uncover hidden patterns, identify nuanced segments, and provide deeper insights into consumer preferences [9]. Deep learning, a subset of machine learning, excels at analyzing large and complex datasets to identify non-linear relationships among variables [3]. Since e-commerce platforms generate vast customer data, deep learning models can integrate diverse information sources, from transactional data and website interactions to textual reviews, to create comprehensive consumer profiles.

Furthermore, leveraging these advanced models allows businesses to make more informed decisions regarding resource allocation and marketing strategies by modeling customer behavior over time. This starkly contrasts traditional methods that often rely on static snapshots of customer interactions. By focusing on CLV prediction through sophisticated machine-learning techniques, e-commerce businesses can enhance their understanding of customer dynamics and optimize their strategies for sustainable growth [8].

Recent industry trends highlight the growing importance of advanced segmentation techniques powered by artificial intelligence and machine learning. Companies are moving beyond static models to dynamic CLV predictions that adapt to real-time consumer behavior. For instance, businesses now use AI to analyze micro-patterns such as cart abandonment rates and engagement times, allowing them to update CLV metrics dynamically and tailor marketing strategies accordingly. Furthermore, predictive analytics enables e-commerce platforms to identify high-value customer segments early and implement targeted campaigns that boost loyalty and spending. Reports suggest that value-based segmentation, when combined with AI-driven insights, has increased CLV by 25% for some businesses.

Specific use cases illustrate how segmentation enhances marketing precision. In the baby products sector, customers purchasing toys exhibit different patterns than those buying consumables like diapers. While toy buyers may remain active for several years, diaper buyers typically have shorter lifespans as customers. By segmenting these groups appropriately, businesses can tailor their strategies for example, offering loyalty programs for toy buyers while focusing on subscription models for diaper customers to maximize CLV in each segment.

This study contains three problem statements of three objectives. The problem statements of this study are: 1) Customer Heterogeneity: E-commerce businesses face the challenge of customer heterogeneity, where diverse behaviors, preferences, and purchasing patterns among customers make it difficult to implement effective marketing strategies. Traditional one-size-fits-all approaches often fail to address this variability, resulting in inefficient resource allocation and missed engagement opportunities; 2) Complexity of Data Analysis: The vast amounts of data generated by e-commerce platforms present significant challenges for businesses. Companies must analyze diverse data sources, including transaction histories and browsing behaviors, and without proper segmentation, critical patterns may be overlooked, limiting the ability to derive actionable insights; and 3) Inaccurate CLV Predictions: Accurately predicting CLV is vital for understanding long-term customer profitability. However, many organizations rely on simplistic models that do not account for the complexities of customer behavior over time, leading to inaccurate predictions and potentially misguided strategic decisions.

The objectives of this study are: 1) to implement customer segmentation based on the characteristics and customer behaviors using clustering techniques which is k-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Models (GMM); 2) to design a machine learning model to predict the CLV of e-commerce customers based on customer information and spending patterns using classification models; and 3) to evaluate the effectiveness of which machine learning model that is performing better for customer segmentation and prediction.

This paper is organized as follows: Section 2 details the literature review of this study. Section 3 describes the method for this study. Section 4 describes the results and discusses the study. Section 5 presents the conclusions of the study.

1. Literature Review
   1. Understanding Customer Lifetime Value (CLV)

Customer Lifetime Value (CLV) is the total revenue a business can expect from a single customer account throughout the business relationship. The accurate prediction of CLV is essential for informing marketing strategies and resource allocation. Kumar et al. [6] emphasize that effective CLV measurement aids in making informed decisions regarding customer acquisition and retention investments, ultimately leading to improved marketing strategies with positive returns on investment (ROI) [6]. Moreover, the significance of CLV extends beyond financial metrics; it influences customer relationship management (CRM) practices, enabling businesses to foster long-term relationships with high-value customers.

* 1. Machine Learning Techniques for CLV Prediction

Recent advancements in machine learning have significantly enhanced the accuracy of CLV predictions. A study by Chen [4] demonstrated that machine learning models outperform traditional methods in predicting CLV, leading to better resource allocation and improved CRM strategies. Various algorithms, including Random Forest, Decision Trees, and Neural Networks, have been employed to analyze customer data effectively. These algorithms allow businesses to uncover complex patterns in customer behavior that traditional statistical methods may overlook. Yashaswini and Prabhudeva [10] explored the Beta-Geometric/Negative Binomial Distribution Model (BG/NBD) and Gamma-Gamma models for predicting customer purchase frequency and value. Their findings indicate that these models can provide detailed insights into customer behavior, facilitating targeted marketing efforts.

* 1. Customer Segmentation Strategies

Segmentation based on CLV is essential for tailoring marketing strategies to different customer groups. The literature reveals various approaches to customer segmentation using machine learning techniques. Abidar et al. [1] utilized cognitive analytics and artificial neural networks to classify customers into high-value segments, enabling firms to prioritize their marketing efforts effectively [1]. This targeted approach enhances marketing efficiency and customer satisfaction by delivering personalized experiences. A study by Shopenova [2] focused on feature engineering and clustering algorithms like K-Means to identify distinct customer segments based on their purchasing behaviour. This approach allows businesses to customize their marketing strategies according to each segment's characteristics [2].

* 1. Multi-Output Models for Enhanced Prediction

Recent research has suggested that incorporating multiple output features alongside CLV can improve prediction accuracy. A study introduced a multi-output deep neural network model that predicts CLV and additional metrics, such as purchase frequency and average transaction size. This model provides a comprehensive view of customer value and aids in more effective segmentation. Predicting multiple dimensions of customer behaviour allows businesses to refine their marketing strategies further, focusing on maximizing overall customer value rather than just individual transactions. Integrating Explainable Artificial Intelligence (XAI) techniques in these models enhances interpretability, allowing businesses to understand better the factors influencing customer value predictions. This transparency is crucial for building trust in machine learning systems within e-commerce contexts.

* 1. Methods comparative analysis

Customer segmentation is a critical component of understanding and engaging with diverse customer bases, and various clustering algorithms are employed to achieve this goal. Three prominent algorithms used in this context are K-means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Gaussian Mixture Model (GMM).

K-means is one of the most widely used clustering algorithms, particularly suited for customer segmentation. It partitions data into a predefined number of clusters (k), with each cluster represented by its centroid. This method is distance-based, grouping data points by minimizing the variance within each cluster. The advantages of K-means include its simplicity and efficiency, making it ideal for large datasets, as well as its interpretability, where each cluster's centroid acts as a "profile" for each customer segment. K-means is the most frequently employed algorithm in customer segmentation research, as highlighted in a systematic literature review that found it used in 20.1% of the reviewed studies [11]. However, K-means assumes spherical clusters and requires a fixed number of clusters to be specified beforehand, which can be challenging without domain knowledge or additional methods like the Elbow Method. Additionally, K-means is sensitive to outliers, as they can skew the clustering results by influencing the centroid calculation.

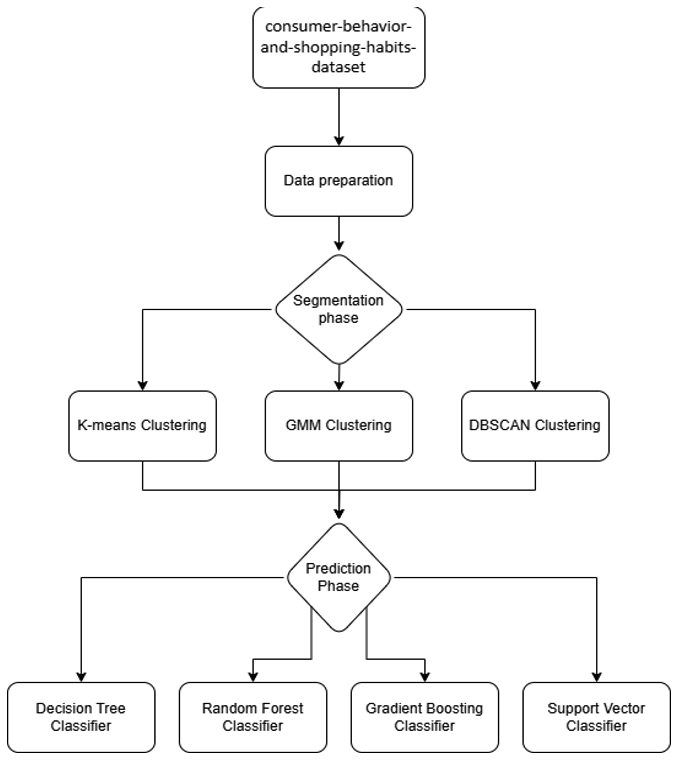
DBSCAN is a density-based clustering algorithm that groups data points based on their density, without requiring a predefined number of clusters. This makes it highly effective for exploratory clustering and detecting clusters of arbitrary shapes. DBSCAN also identifies outliers as noise, which is useful for pinpointing atypical customers. The advantages of DBSCAN include its ability to find clusters without specifying their number and its effectiveness in detecting irregularly shaped clusters, which may better represent real-world customer behaviours. However, DBSCAN is sensitive to parameter selection, particularly the eps and min\_samples parameters, which can significantly impact clustering results. Additionally, it struggles with data containing clusters of differing densities, as it assumes a consistent density within clusters [11]

GMM is a probabilistic model that assumes data is generated from a mixture of several Gaussian distributions, each representing a cluster. This model is more flexible than K-means, allowing clusters to have different shapes and modeling overlapping clusters effectively. GMM provides a probabilistic assignment, meaning each point can belong to multiple clusters with certain probabilities, making it effective for overlapping clusters. It also accommodates elliptical clusters with varying sizes and orientations, making it more adaptable to complex data structures. However, GMM requires specifying the number of components in advance, like K-means, and is computationally intensive due to the iterative optimization of multiple parameters for each Gaussian component. Despite these challenges, GMM offers a powerful tool for capturing nuanced customer behaviors and preferences.

In addition to clustering algorithms, classification algorithms like Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier, and Support Vector Classifier are also used for customer segmentation. The Decision Tree Classifier offers a straightforward and interpretable approach but tends to overfit, necessitating techniques like pruning to improve generalization. The Random Forest Classifier is robust and reduces overfitting by creating an ensemble of decision trees. It provides estimates of feature importance, which is crucial for identifying predictive customer attributes. Its proven accuracy in e-commerce customer segmentation, such as achieving 75.50% accuracy in a related project, further supports its choice [12]. Gradient Boosting Classifier is chosen for its high predictive power and ability to capture complex relationships in the data. It often achieves state-of-the-art performance but requires careful hyperparameter tuning to avoid overfitting. Its success in achieving 75.77% accuracy in another e-commerce segmentation project highlights its potential [12]. Support Vector Classifier is effective in high-dimensional spaces and can handle non-linear relationships using kernel functions. However, its computational expense and sensitivity to hyperparameters necessitate careful optimization [13].

1. Methodology

Fig. 1 shows the methodology of the study. The dataset must go through data preparation where it needs to be clean, normalized, understood, and be ready for the next phase, which is the segmentation phase, where K-means, GMM, and DBSCAN are used to segment the customer into low, medium or high value because the dataset does not have the label. Next, the customer value segmentation label from the segmentation phase can be used as a target variable in the prediction phase, where a decision tree classifier, random forest classifier, gradient boosting classifier, and support vector classifier are used to implement prediction on each clustering model.



**Fig. 1.** Flowchart of the overall experiment design

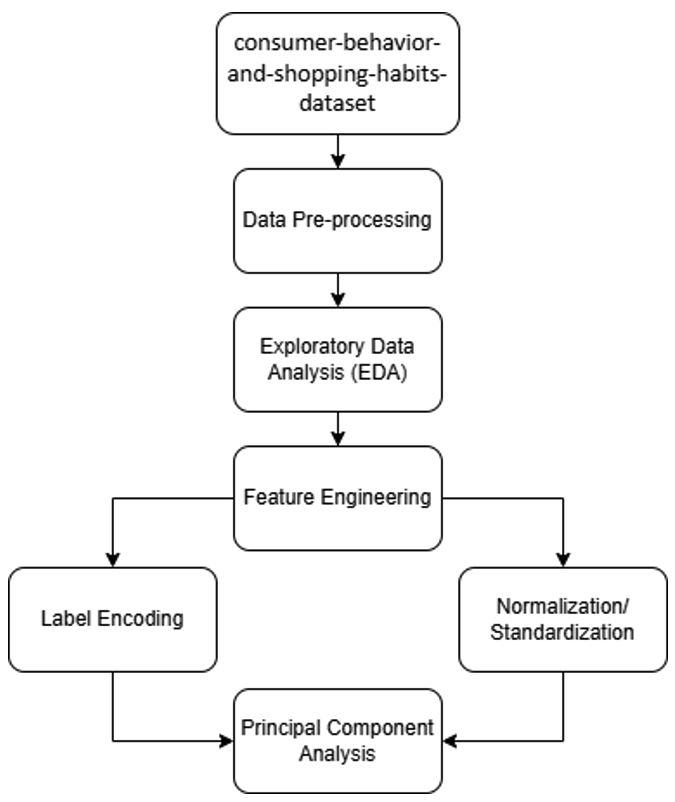
* 1. Data Collection and Preparation

The dataset was obtained from Kaggle, where the owner collects the data from unknown e-commerce platforms because the owner does not want to leak the e-commerce due to the e-commerce company's privacy. Table 1 shows the details of each feature in the dataset; these features were used later for feature engineering.

**Table 1.** List of features in the dataset

|  |
| --- |
| Features |
| Customer ID |
| Age |
| Gender |
| Item Purchased and Category |
| Purchase Amount (USD) |
| Location |
| Size and Color |
| Season |
| Review Rating |
| Subscription Status |
| Shipping Type |
| Discount Applied and Promo Code Used |
| Previous Purchases |
| Payment Method |
| Frequency of Purchases |

Fig. 2 shows the data preparation process. The flow chart begins with the dataset, which consists of the raw data collected for customer segmentation. This data may contain inconsistencies, missing values, or irrelevant information that must be addressed for accurate analysis. The pre-processing stage involves two sub-processes: detecting and clearing missing values and identifying and removing duplicate rows. Following this, Exploratory Data Analysis (EDA) is conducted, generating summary statistics and visualizations, such as histograms and scatter plots, to uncover insights into customer behavior and guide feature engineering and model selection. Finally, in the Feature Engineering step, derived features like Total Spend, Loyalty Score, Customer Characteristics, Purchase Behavior, and Composite Score are created to capture key customer behaviors and enhance the dataset's value. See Table 2 for the details of the new features.



**Fig. 2.** Flowchart of the data preparation process

**Table 2.** Derived features formula

|  |  |
| --- | --- |
| Derived Features | Formula |
| Total Spend | Calculates the cumulative spending for each customer by multiplying the number of previous purchases with the purchase amount of the current transaction. |
| Customer Characteristics | Combines multiple customer attributes (specifically, age, gender, and location) into a single feature. |
| Purchase Behaviour | Combines multiple attributes related to each customer’s purchase preferences into a single feature (size, colour, season, category, review writing). |
| Loyalty Score | Custom loyalty scoring system for each customer, using a range of factors that indicate loyalty, such as subscription status, frequency of purchases, and usage of discounts and promo codes. |
| Composite Score | Calculated as a weighted sum of total spend, customer characteristics, purchase behaviour and loyalty score, with each feature contributing a certain percentage to theoverall score |

These engineered features can significantly improve the quality of customer segmentation by incorporating domain knowledge and adding higher-level insights about customer patterns. Standardization or normalization and label encoding are performed to ensure that each feature contributes equally to clustering, preventing features with large numeric ranges from dominating the analysis. In addition, Principal Component Analysis (PCA) is applied for dimensionality reduction, which helps visualize the clusters and simplifies the complex feature while retaining the most significant variance in the dataset.

Data preparation is a crucial step in the analysis process, ensuring that the data is clean and suitable for further analysis or model development. To achieve this, data mining techniques are applied to the raw dataset to preprocess and clean the data, making it correct, consistent, and ready for analysis. Data cleaning and preprocessing are essential to ensure the dataset is free from errors, inconsistencies, and irrelevant information. Techniques such as missing value handling—where missing values are replaced with the mean, median, or mode, or rows with missing values are removed—and duplicate detection and removal are used to enhance data quality. This step is vital because it directly impacts the accuracy and reliability of the results obtained from data mining algorithms.

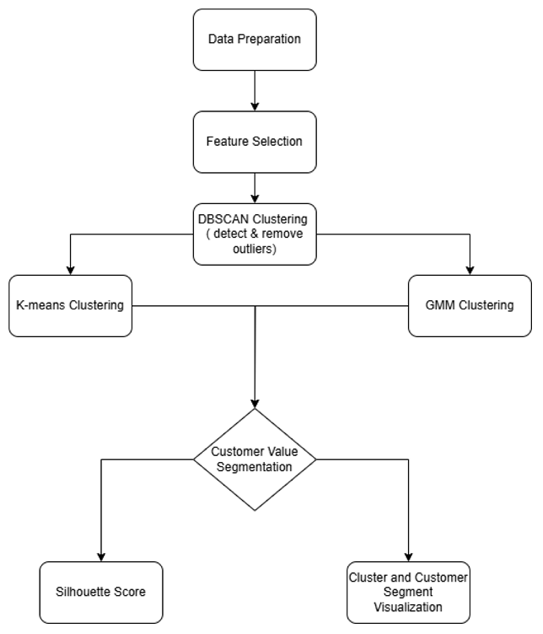
Data transformation, including standardization and normalization, is another critical step. Standardization scales feature to have a mean of 0 and a standard deviation of 1, maintaining the distribution shape but adjusting the scale, which is beneficial for algorithms sensitive to feature magnitude. Normalization rescales features to a common range, typically between 0 and 1, ensuring that all features contribute equally to the clustering process. This prevents features with larger ranges from dominating the results, ensuring a balanced analysis.

Feature engineering plays a pivotal role in making the dataset more informative and suitable for clustering. This involves creating new features or modifying existing ones, such as deriving features like Total Spend, Loyalty Score, or Customer Characteristics, and encoding categorical features using methods like Label Encoding or One-Hot Encoding. Additionally, aggregation and transformation techniques are used to capture customer behaviours, such as summarizing purchase frequency or average spending. By enhancing the dataset in this way, feature engineering improves the accuracy and interpretability of the segmentation results, allowing for more meaningful patterns to be revealed.

* 1. Segmentation Phase

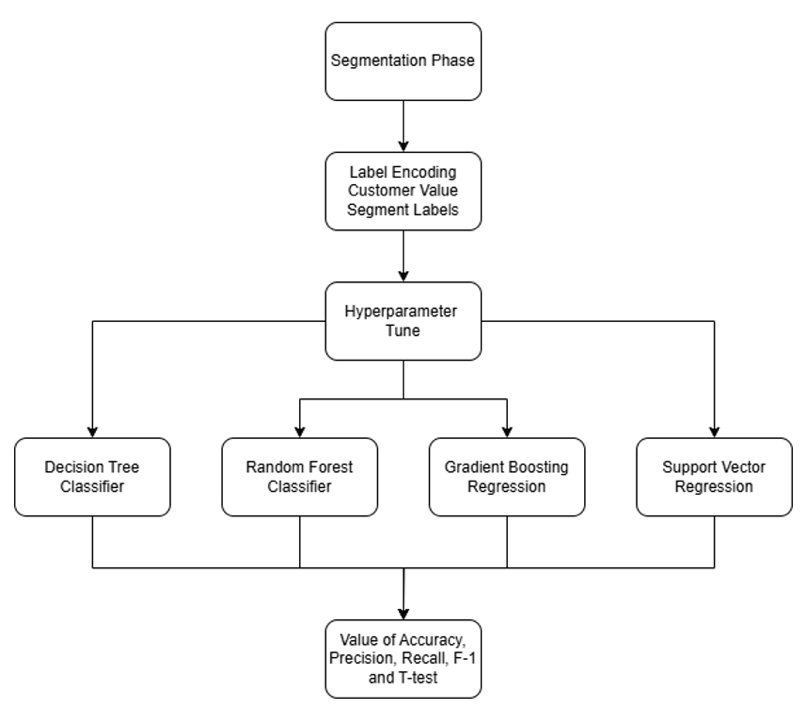
Fig. 3 shows the process of the data segmentation phase. The process begins with feature selection, where the most relevant variables are identified to contribute to the clustering model and create meaningful customer segments. DBSCAN clustering is then performed to identify clusters of dense points while labeling sparse points as noise (outliers). These outliers are removed from the dataset to ensure cleaner data for subsequent clustering steps. Once the dataset is refined, K-means clustering is applied using selected features, such as Composite Score and Total Spend, along with optimal hyperparameters to segment customers into distinct clusters. Similarly, the Gaussian Mixture Model (GMM) is employed with the same features and an optimal number of components to further segment customers.

The Silhouette Score is chosen as a key measurement metric for evaluating the performance of clustering models due to its ability to assess both the cohesion within clusters and the separation between them. This metric provides valuable insights into how well each data point fits within its assigned cluster, which is crucial for ensuring that customer segments are well-defined and distinct. By comparing Silhouette Scores across different clustering models or configurations, researchers can select the model that best captures the underlying structure of the customer data. This metric aids in optimizing model parameters to achieve the most coherent and distinct customer segments. Finally, the results are visualized using scatter plots, bar charts, or other visual tools, with each cluster represented by a unique color. These visualizations help illustrate and interpret the distinct customer segments created by the clustering models.



**Fig. 3.** Flowchart of the segmentation phase

* 1. Prediction Phase



**Fig. 4.** Flowchart of the prediction phase

Fig. 4 shows the prediction phase process. Label encoding and normalization of customer value segment labels are essential to preparing data for machine learning. Label encoding converts categorical variables into numerical format by assigning distinct integers to each unique segment label, while normalization scales these values to a common range, typically between 0 and 1, to prevent any single label from disproportionately influencing the model. This process is crucial for maintaining data integrity and enhancing model performance. In addition, the train-test split ratio, often set at 80:20, allocates 80% of the dataset for training the model and 20% for testing its performance. This division helps assess how well the model generalizes to unseen data, thus preventing overfitting. Finally, hyperparameter tuning optimizes the parameters that control the learning process of machine learning models but are not learned from the data itself, further improving model accuracy and effectiveness. See Table 3 for the list of the hyperparameter tunes.

**Table 3.** List of hyperparameter tunes for each classification model

|  |  |
| --- | --- |
| Classification Model | Hyperparameter Tunes |
| Decision tree | Max\_depth: 3, 5, 10, None |
|  | Min\_samples\_split: 2, 5, 10 |
| Random forest | N\_estimators: 50, 100, 200 |
|  | Max\_depth: 3, 5, None |
| Gradient boosting | N\_estimators: 50, 100, 200 |
|  | Learning\_rate: 0.01, 0.1, 0.2 |
| Support vector (SVM) | C: 0.1, 1, 10 |
|  | Kernel: linear, rbf |

Multiple classification models are implemented to predict customer segments based on dataset-derived features, including the Decision Tree Classifier, a simple yet powerful algorithm that splits data based on feature conditions. The Random Forest Classifier, an ensemble method, constructs multiple decision trees and uses the majority vote for final classification to improve accuracy and reduce overfitting. The Gradient Boosting Classifier, a boosting technique, builds models sequentially, correcting errors from previous models to enhance performance on complex data.

The Support Vector Classifier (SVC) also finds the optimal hyperplane to separate classes, ensuring maximum margin between customer segments. To evaluate the performance of each model, metrics such as Accuracy (the percentage of correctly classified segments), Precision (the proportion of correctly predicted positive instances), Recall (the proportion of correctly predicted positive instances out of all actual positive instances), and the F1-Score (the harmonic means of precision and recall) are used. Finally, a Paired T-Test is conducted to compare model performance by checking whether accuracy differences between models are statistically significant.

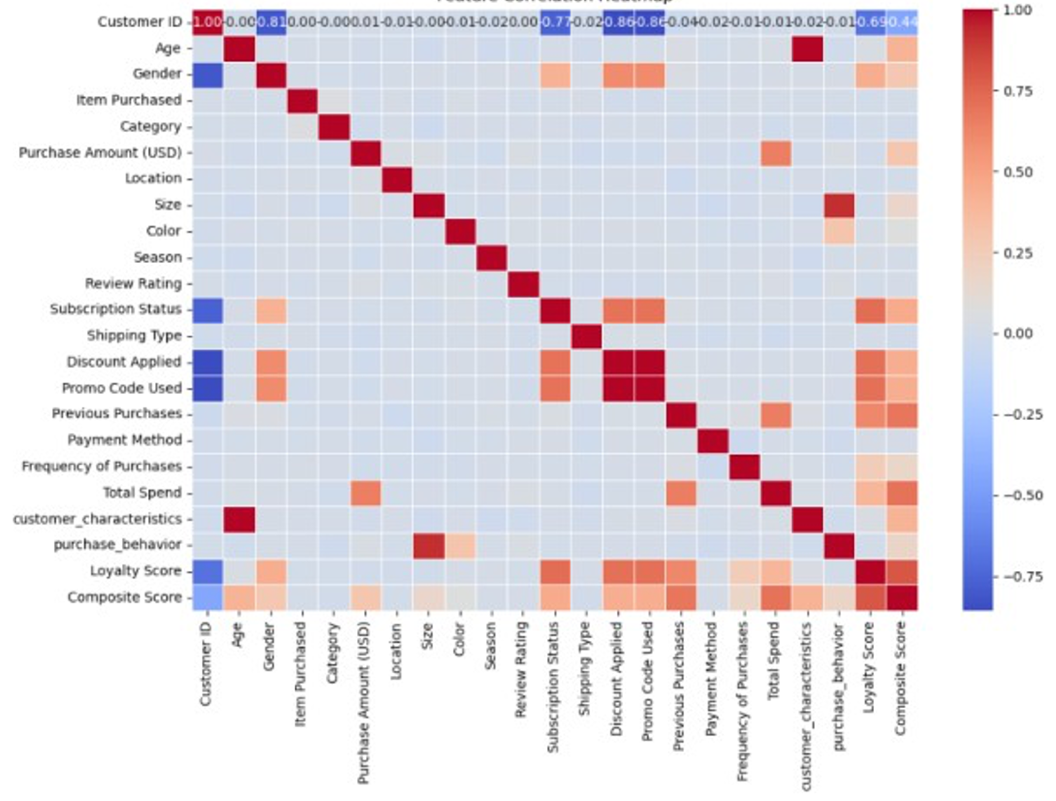
The choice of metrics for evaluating classification models is driven by their ability to provide a comprehensive assessment of model performance. Accuracy offers a general overview by measuring the proportion of correct predictions, although it can be misleading in imbalanced datasets. Precision is critical in scenarios where the cost of false positives is high, such as in targeted marketing, ensuring that predicted high-value customers are indeed likely to be so. Recall is essential for identifying all relevant instances, particularly important when missing a positive instance has significant consequences, like failing to identify high-value customers at risk of churn. The F1-score balances precision and recall, providing a single metric that is particularly useful in imbalanced datasets where one class may be more important than another. Additionally, the T-test, while not a direct classification metric, compares the means of different groups, offering insights into whether customer segments identified by different models have statistically significant differences in Customer Lifetime Value (CLV). This combination of metrics ensures that model performance is evaluated from multiple angles, providing a robust foundation for decision-making in customer segmentation.

These metrics collectively provide a comprehensive view of model performance, addressing both the accuracy of predictions and the balance between precision and recall.Metrics like Precision, Recall, and F1-score are crucial for evaluating models in datasets where one class significantly outnumbers another, ensuring that the model's performance is not skewed by the dominant class. The T-test adds a layer of statistical validation, ensuring that the differences between customer segments are not only practically significant but also statistically robust.

1. Results and Discussions

This section presents the results and discusses segmentation, prediction, analysis, and model evaluation results.

The feature correlation heatmap (see Fig. 5) illustrates the relationships between different dataset features. Each cell shows the correlation coefficient, ranging from -1 to 1. A red color indicates a strong positive correlation (close to 1), meaning both features increase together, while blue indicates a strong negative correlation (close to -1), showing that feature decreases as the other increases. Colors near white suggest weak or no correlation.



**Fig. 5.** Feature correlation heatmap

'Composite Score' strongly correlates positively with 'Loyalty Score' and 'Total Spend', but negatively with 'Subscription Status', 'Discount Applied', and 'Promo Code Used'. These insights into feature dependencies aid in feature selection and model building. Consequently, 'Composite Score' and 'Total Spend' were chosen as target variables for segmentation or clustering.

* 1. Clustering

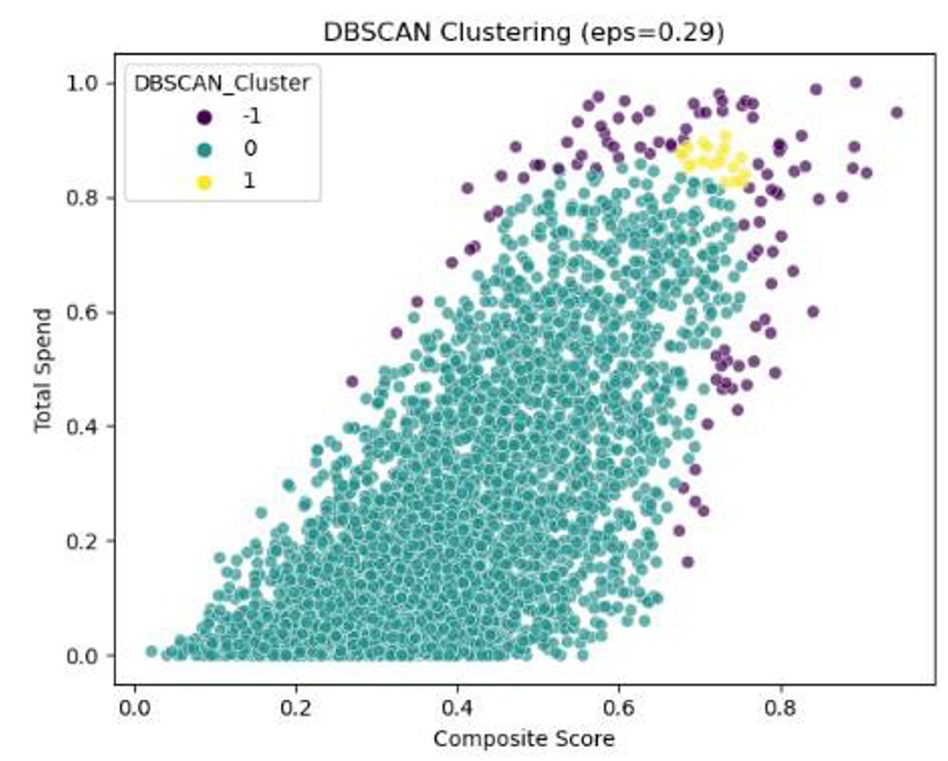
Before performing the clustering techniques, it is important to find the best value of K to optimize the clustering performance. The elbow method graph (see Fig. 6) shows K=3 appears to be a reasonable choice for the best number of clusters. The Elbow Method involves plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters (K). The "elbow" point on the graph is where the rate of decrease in WCSS starts to diminish significantly. The drop in WCSS from K=3 to K=4 is less substantial than from K=2 to K=3. Therefore, K=3 is the better choice.

A graph with a blue line

AI-generated content may be incorrect.

**Fig. 6.** Graph for elbow methods to choose the K-value

Fig. 7 shows the scatter plot that visualizes the results of DBSCAN clustering on a dataset, with each point representing a customer plotted according to their "Composite Score" (x-axis) and "Total Spend" (y-axis). The plot uses color to indicate cluster membership as determined by DBSCAN, with the parameter eps (epsilon, the maximum distance between two samples for them to be considered in the same neighbourhood) set to 0.29.



**Fig. 7.** Scatter plot of DBSCAN clustering

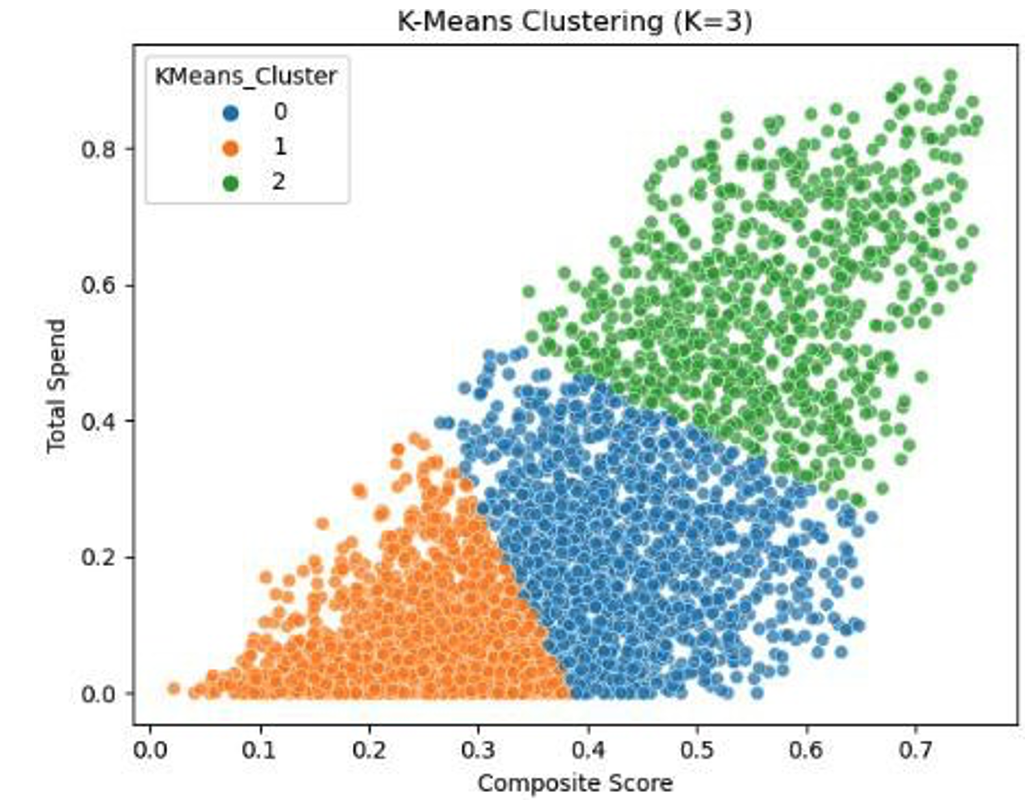
Green points (Cluster 0): Form the large, dense cluster in the bottom-left to the middle-right of the plot. These represent most customers, who have a wide range of total spending and composite scores.

Purple points (Cluster -1): Represent points labeled as noise or outliers by DBSCAN. These customers do not belong to any dense cluster based on the defined eps value.

Yellow points (Cluster 1): Form a small, dense cluster in the upper-middle-right of the plot. These customers have high composite score, and total spend.

The DBSCAN algorithm has identified a large group of customers with a wide range of total spend and composite score, with an eps of 0.29. It also identifies outliers and customers with high total spend and composite scores.

Fig. 8 shows the scatter plot of K-Means clustering (with K=3) applied to a dataset, plotting each customer based on their "Composite Score" (x-axis) and "Total Spend" (y-axis). The plot uses different colors to represent the three clusters identified by the algorithm.



**Fig. 8.** Scatter plot of K-means clustering

Cluster 0 (Blue): This cluster appears to represent customers with a moderate composite score (ranging from about 0.35 to 0.7) and a moderate total spend.

Cluster 1 (Orange): This cluster is formed by customers with low composite score (ranging from 0 to 0.35) and low total spend (0 to 0.3). These may represent new customers who haven't spent a lot of money yet.

Cluster 2 (Green): This cluster seems to represent customers with high composite score and high total spend

K-Means has segmented the customer base into three distinct groups based on their "Composite Score" and "Total Spend": low-spending, low-score customers (orange), moderate-spending, moderate-score customers (blue), and high-spending, high-score customers (green).

Fig. 9 shows the scatter plot of GMM clustering with 3 components (n=3), plotting each customer by "Composite Score" (x-axis) and "Total Spend" (y-axis). Different colors represent the clusters identified by the algorithm.

A diagram of a clustering graph

AI-generated content may be incorrect.

**Fig. 9.** Scatter plot of GMM clustering

Cluster 0 (Blue): These customers generally have moderate "Composite Scores" (ranging from approximately 0.2 to 0.6) and moderate "Total Spend" (0 to 0.3).

Cluster 1 (Gray): This cluster consists of customers with higher "Composite Scores" (ranging from 0.3 to 0.75) and generally higher "Total Spend" (0.4 to 0.8).

Cluster 2 (Red): This cluster represents customers with low total spending and low composite scores.

The GMM algorithm has segmented the customer data into three distinct groups: low, moderate, and high-value customers, as defined by their Composite Score and Total Spend. The grey segments in the plot have a positive trend; that is, the higher total spend will have a higher composite score. The red segments have low spending and low scores. The blue segments have a moderate score with moderate spending.

Table 4 shows the silhouette score comparison; DBSCAN achieves the highest silhouette score among the three models, indicating that it has formed well-defined clusters with clear boundaries. K-Means achieves a moderate silhouette score, slightly lower than DBSCAN's. This shows that while K- Means forms reasonably distinct clusters, there may be some overlap or less well-defined boundaries between clusters. GMM has the lowest silhouette score, indicating that its clusters are less distinct, and more overlapping compared to DBSCAN and K-Means.

**Table 4.** Silhouette score comparison between clustering models

|  |  |
| --- | --- |
| Clustering model | Score |
| DBSCAN | 0.48 |
| K-means (k=3) | 0.41 |
| GMM (n=3) | 0.26 |

* 1. Classification

Table 5 shows the classification models; Decision Tree, Random Forest, and Gradient Boosting; consistently achieve high accuracy, ranging from 97% to 99%, across both clustering methods. In contrast, SVM underperforms in both segmentation approaches and exhibits a higher standard deviation, indicating less stability. When comparing segmentation methods, GMM segmentation results in slightly better and more stable model performances than K-Means. Among the classifiers, Gradient Boosting is the best model in K-Means segmentation, while Decision Tree and Gradient Boosting perform equally well in GMM segmentation. Random Forest is highly stable in GMM but shows slightly weaker performance in K-Means.

**Table 5.** Performance comparison of each classification model by using mean accuracy and standard deviation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Mean Accuracy  (k- means) | Std Dev  (k- means) | Mean Accuracy  (k- means) | Std Dev  (k- means) |
| Decision tree | 0.9773 | 0.0066 | 0.9911 | 0.0059 |
| Random forest | 0.9691 | 0.0088 | 0.9908 | 0.0034 |
| Gradient boosting | 0.9845 | 0.0025 | 0.9911 | 0.0057 |
| SVM | 0.9589 | 0.0069 | 0.9381 | 0.0112 |

Based on these findings, Gradient Boosting is the best overall model, demonstrating consistently high accuracy and stability in both segmentation techniques. It also slightly outperforms other models when using K-Means segmentation. As alternative strong models, Decision Tree and Random Forest are viable options. Decision Tree performs equally well as Gradient Boosting in GMM segmentation, while Random Forest is the most stable model in GMM, though slightly weaker in K-Means. On the other hand, SVM is not recommended due to its lower accuracy and higher variance across both clustering techniques. It is significantly weaker than ensemble-based models such as Decision Tree, Random Forest, and Gradient Boosting and does not offer competitive performance in this classification task.

* 1. Discussion on model performance and interpretation

The results indicate that different classification models perform variably depending on the segmentation approach used. Gradient Boosting consistently emerges as the best overall model, particularly excelling in K-Means segmentation, while Decision Tree performs equally well under GMM segmentation. Random Forest, though highly stable, shows slightly weaker performance in K-Means. In contrast, SVM consistently underperforms in both segmentation methods, highlighting its limitations in this context.

A deeper analysis of why certain models outperformed others reveals key insights into the nature of the data and the suitability of different algorithms:

Gradient Boosting, which sequentially improves predictions by minimizing residual errors, achieves high accuracy due to its ability to capture complex relationships within the dataset. Decision Trees, despite their simplicity, perform well in GMM segmentation, likely due to the segmentation’s ability to structure data in a way that aligns well with tree-based splitting criteria. Random Forest, as an ensemble of decision trees, provides stability and reduces overfitting but lacks the adaptive boosting mechanism of Gradient Boosting, which may explain its slightly lower performance in K-Means segmentation.

K-Means clustering provides well-defined, distinct customer groups, benefiting models like Gradient Boosting that rely on clear boundaries between classes. MM, which assumes Gaussian distributions for clusters, results in some overlap between customer groups. This may explain why Decision Tree performs equally well as Gradient Boosting, as it effectively captures decision boundaries in slightly less distinct clusters.

The reliance on silhouette scores for clustering evaluation suggests that cluster compactness was prioritized, but real-world customer segmentation may not always be well-defined. The models might be sensitive to imbalanced class distributions. If one segment has significantly fewer customers, models may have difficulty learning generalizable patterns for that segment. SVM’s lower accuracy and higher variance suggest that a linear separation assumption does not hold for this dataset, and its performance is further impacted by the nature of the feature space.

The integration of clustering with classification models enhances predictive performance compared to using classification alone. The best strategy for customer lifetime value prediction involves using DBSCAN for outlier detection, K-Means for segmentation, and Gradient Boosting for classification, ensuring both accuracy and stability in predictions. Further analysis on potential biases in segmentation and feature engineering could enhance model interpretability and real-world applicability.

The classification models used in this study—Gradient Boosting, Decision Tree, Random Forest, and Support Vector Machine (SVM)—demonstrated varying levels of performance based on the chosen segmentation approach. Overall, Gradient Boosting emerged as the best-performing model, particularly excelling in K-Means segmentation. Its ability to sequentially refine weak learners contributed to its high accuracy and strong generalization ability. Decision Trees also performed well, especially in GMM segmentation, where their rule-based splits aligned effectively with the cluster structure. Meanwhile, Random Forest exhibited high stability across both segmentation methods but showed slightly lower performance in K-Means compared to Gradient Boosting.

SVM, on the other hand, consistently underperformed across both segmentation techniques. Its reliance on finding an optimal hyperplane proved ineffective due to the complex and non-linearly separable nature of customer data. The RBF kernel struggled to capture meaningful patterns, leading to lower accuracy and higher variance. This suggests that tree-based models are more suitable for customer segmentation, as they can better handle overlapping distributions and nonlinear relationships within the data.

The choice of segmentation method significantly influenced classification performance. K-Means clustering, which creates well-defined, distinct groups, allowed models to achieve clearer decision boundaries, benefiting Gradient Boosting and Decision Tree the most. In contrast, GMM segmentation resulted in overlapping clusters, which made classification more challenging. Interestingly, Decision Tree performed equally well as Gradient Boosting in this scenario, likely due to its ability to partition feature space efficiently. However, the slight drop in Gradient Boosting’s performance in GMM suggests that it may struggle with segmentations that lack clear separations.

Despite strong results, certain limitations and biases should be considered. The presence of imbalanced customer segments may have affected classification accuracy, potentially leading models to favor dominant classes. Additionally, while Gradient Boosting achieved the highest accuracy, its black-box nature makes interpretation more difficult compared to Decision Trees and Random Forest, which offer clearer insights into feature importance. Overfitting remains another concern, particularly for ensemble models with complex learning processes, highlighting the need for careful hyperparameter tuning and validation on unseen data.

In conclusion, the study confirms that ensemble-based models, particularly Gradient Boosting, are well-suited for customer lifetime value prediction. Decision Tree and Random Forest serve as strong alternatives, especially when model interpretability or stability is a priority. The findings also emphasize that K-Means segmentation provides clearer customer groups, leading to better classification results. Future work could explore hybrid approaches that combine multiple classifiers or further refine feature selection to enhance model interpretability and real-world applicability.

1. Conclusion

The findings of this study offer valuable insights for e-commerce businesses looking to optimize customer segmentation and predictive modelling strategies. By leveraging Gradient Boosting for classification and K-Means for segmentation, businesses can more accurately identify high-value customers, predict future purchasing behaviours, and tailor marketing strategies accordingly. This data-driven approach enhances personalization efforts, allowing businesses to maximize customer retention and lifetime value.

One key application is targeted marketing campaigns. With clear segmentation of customers into low, medium, and high-value groups, businesses can create personalized promotions for each segment. For instance, high-value customers—those with strong purchasing patterns—can be incentivized with exclusive loyalty rewards or early access to new products, fostering brand loyalty. Meanwhile, medium-value customers can receive targeted discount offers or personalized recommendations to encourage repeat purchases. Low-value customers, who may be at risk of disengagement, can be re-engaged with special discounts, retargeting ads, or automated follow-up emails to nurture brand interaction.

Beyond marketing, these insights can also inform inventory management and dynamic pricing strategies. By understanding purchasing behaviours across different customer segments, businesses can optimize stock levels for popular products, reducing overstock and minimizing lost sales due to inventory shortages. Additionally, predictive modelling can support personalized pricing strategies, offering discounts or bundling options based on a customer’s likelihood of purchasing specific products.

Integrating these predictive analytics models into a customer relationship management (CRM) system can enhance automated decision-making. Businesses can use real-time data to adjust marketing strategies dynamically, optimize advertising spend, and improve overall customer experience. By adopting these AI-driven approaches, e-commerce businesses can create more effective, data-backed strategies that drive long-term profitability and customer satisfaction.

This study successfully achieved its objectives through a structured process involving data preparation, segmentation, prediction, and evaluation. In the data preparation phase, feature engineering enhanced segmentation effectiveness by incorporating a comprehensive set of independent variables. K-Means, GMM, and DBSCAN, were explored during the segmentation phase to categorize customers into low, medium, and high-value groups. The results showed that K-Means balanced simplicity and effectiveness, while DBSCAN was particularly useful for detecting and removing outliers. GMM, however, struggled with overlapping clusters, making it less effective for segmentation.

Four classification models, Decision Tree, Random Forest, Gradient Boosting, and SVM, were evaluated in the prediction phase. Gradient Boosting emerged as the best-performing model, delivering high accuracy and stability across different segmentation methods. Decision Tree and Random Forest also performed well, particularly in GMM-based segmentation, while SVM showed weaker performance with lower accuracy and higher variance. Hyperparameter tuning was applied to each model to optimize performance, and evaluations were conducted using accuracy, precision, recall, F1-score, and T-tests to ensure robustness. Based on the results, the recommended approach is to use DBSCAN for outlier detection, K-Means for segmentation, and Gradient Boosting for classification to achieve the most accurate and stable customer lifetime value prediction.

While the results are promising, future work should validate these findings using a separate test dataset and incorporate additional evaluation metrics to assess model reliability and generalizability further. In addition, integrating external data sources such as economic indicators or competitive analysis could enhance the dataset’s richness, improving predictive accuracy and model robustness. This approach could provide deeper insights into customer behavior, refining CLV predictions and decision-making processes.

References

1. M. Abidar, S. Alshahrani, and M. Alzahrani, "Cognitive analytics for customer segmentation: Enhancing marketing strategies through machine learning techniques," *J. Bus. Res.*, vol. 145, no. 1, pp. 234- 245, 2023.
2. A. Shopenova, "Feature engineering techniques for customer segmentation: Clustering approaches in e-commerce contexts," *Int. J. Inf. Syst.*, vol. 9, no. 2, pp. 67-79, 2021.I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
3. N. Ali, U. Awan, and M. Khan, "Customer lifetime value (CLV) insights for strategic marketing practices: An empirical analysis," *J. Mark. Anal.*, vol. 10, no. 1, pp. 1-15, 2024. doi: 10.1080/23311975.2024.2361321.
4. Y. Chen, "Machine learning approaches for customer lifetime value prediction: A comparative study," *J. Mark. Anal.*, vol. 6, no. 3, pp. 150- 162, 2018.
5. K. Donnelly, "The importance of customer lifetime value in eCommerce," Invesp, 2023. [Online]. Available: [https://www.invespcro.com/blog/the-importance-of-customer-](https://www.invespcro.com/blog/the-importance-of-customer-lifetime-value-in-ecommerce/) [lifetime-value-in-ecommerce/.](https://www.invespcro.com/blog/the-importance-of-customer-lifetime-value-in-ecommerce/)
6. A. Kumar, R. Singh, and P. Gupta, "Measuring customer lifetime value: A comprehensive review of approaches and applications in e- commerce," *Int. J. Comput. Appl.*, vol. 24, no. 4, pp. 12-23, 2023.
7. Z. Tabaei and M. Fathian, "Using customer lifetime value model for product recommendation: An electronic retailing case study," *Int. J. e- Educ., e-Bus., e-Manag. e-Learn.*, vol. 9, no. 1, pp. 77-82, 2019. [Online]. Available: [https://www.ijeeee.org/Papers/084-](https://www.ijeeee.org/Papers/084-Z00067T84.pdf) [Z00067T84.pdf.](https://www.ijeeee.org/Papers/084-Z00067T84.pdf)
8. Verfacto, "Customer lifetime value in e-commerce - why it is important," 2023. [Online]. Available: [https://www.verfacto.com/blog/customer-lifetime-value/in-](https://www.verfacto.com/blog/customer-lifetime-value/in-ecommerce/) [ecommerce/.](https://www.verfacto.com/blog/customer-lifetime-value/in-ecommerce/)
9. J. Wang, Y. Zhang, and H. Liu, "Business analytics in customer lifetime value: A review across industries," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 14, no. 1, e1571, 2024. [Online]. Available: [https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1571.](https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1571)
10. S. Yashaswini and K. Prabhudeva, "Predicting customer purchase behavior using BG/NBD and Gamma-Gamma models: An empirical study in e-commerce," *Int. J. Data Sci.*, vol. 13, no. 7, pp. 45-58, 2022.
11. E. Ayodele and V. Sodeinde, “Customer Segmentation Using the K-Means Clustering Algorithm,” IJST. Accessed: Mar. 03, 2025. [Online]. Available: <https://sciencetechjournal.federalpolyilaro.edu.ng/storage/article/JST%2015%20Done_1729671956.pdf>
12. W. Xuan, U. Tunku, and A. Rahman, “Title Page CUSTOMER SEGMENTATION ON CLUSTERING ALGORITHMS A REPORT SUBMITTED TO.” Accessed: Aug. 21, 2024. [Online]. Available: <http://eprints.utar.edu.my/5523/1/fyp_CS_2023_TWX.pdf>
13. J. Salminen, Mekhail Mustak, Muhammad Sufyan, and B. J. Jansen, “How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation,” *Journal of marketing analytics*, vol. 11, Jul. 2023, doi: https://doi.org/10.1057/s41270-023-00235-5.