# A Study on Enhancing Customer Reactivation through RFM Analysis and Predictive Analytics

Prof (Dr). S.L. Gupta
*Professor, Birla Institute of Technology Noida campus*

***Abstract*—Understanding consumer behavior and making accurate predictions about their propensity to return are essential for increasing customer lifetime value and retention in the cutthroat world of contemporary retail. Despite being a popular technique for determining customer value, the traditional Recency, Frequency, and Monetary (RFM) analysis frequently fails in dynamic markets because it relies on rule-based models and static clustering techniques that do not support customized marketing strategies. These restrictions lead to the treatment of consumer segmentation and reactivation as distinct procedures, which lowers the efficacy of marketing as a whole. This research combines RFM analysis with Random Forest classification and K-Means clustering to provide a data-driven, integrated approach that improves client reactivation efforts. The method uses unsupervised K-Means clustering to first divide up the customer base according to their buying patterns. Then, it uses the cluster characteristics that are produced in a supervised Random Forest model to forecast the chance that inactive customers would reactivate. A more sophisticated and predictive knowledge of consumer dynamics is guaranteed by the model's integration of the advantages of supervised and unsupervised learning. When applied to a real-world retail dataset, the framework demonstrated a great clustering performance, exhibiting meaningful and well-separated clusters with a Calinski-Harabasz Index of 14,913.66 and a Silhouette Score of 0.5524. Furthermore, 97.8% classification accuracy and zero misclassifications were attained by the prediction model on the test set, indicating its practical applicability and resilience. The creation of more focused and successful marketing campaigns is made possible by this thorough technique, which closes the gap between client segmentation and reactivation forecast. Finally, by assisting in the creation of intelligent CRM (Customer Relationship Management) systems that maximize marketing budgets and enhance customer retention through prompt and tailored re-engagement campaigns, the research provides a promising avenue for further research in predictive customer analytics.**

***Keywords*— Customer Reactivation, RFM Analysis, Customer Segmentation, K-Means Clustering, Random Forest Classifier, Predictive Analytics.**

## I. INTRODUCTION

Retaining and reactivating customers has become essential for maintaining business growth and profitability in the modern digital economy. Revenue must put in place efficient tactics that promote repeat business and re-engage dormant Customers because acquiring new customers frequently comes with higher costs than keeping existing ones. The RFM analysis (Recency, Frequency, Monetary) is one of the fundamental models used in consumer segmentation. Three factors are used in this approach to assess a customer's value: the frequency of their purchases, the amount they spend, and the recentness of their purchases (Alet Vilaginés, 2020), as illustrated in Figure 1. Finding high-value Customers and adjusting marketing tactics appropriately have been made possible thanks in large part to RFM analysis.

But there are drawbacks to standard RFM analysis, especially with regard to its static character and incapacity to record intricate Customer behaviors over time (Ho et al., 2023). A more dynamic and perceptive method of consumer segmentation and retention is provided by combining RFM analysis with machine learning approaches, such as time series forecasting and clustering algorithms, to address these issues (Lewaaelhamd, 2023).

The effectiveness of K-Means clustering, an unsupervised machine learning approach, in dividing datasets into discrete groups according to feature similarity is well known (Huang et al., 2021). K- Means clustering can reveal hidden patterns and divide consumers into meaningful groups when applied to RFM variables, enabling more focused marketing campaigns. For example, a study by (Wong et al., 2024) showed how well RFM analysis and K-Means clustering work together to discover Customer categories with unique buying patterns, which results in more specialized marketing tactics. Time series prediction algorithms like Prophet and ARIMA (Auto Regressive

Integrated Moving Average) have shown promise in predicting customer behavior over time in addition to clustering (Liço et al., 2021). Businesses can proactively engage customers before they churn by using these algorithms to forecast future purchasing trends. For instance, (Xiahou & Harada, 2022) modelled customer lifecycle behavior using time series analysis, which helped them understand retention dynamics and guide reactivation tactics.

A thorough framework for comprehending and influencing consumer behavior is produced by combining RFM analysis with K-Means clustering and time series prediction. Businesses can use this integrated strategy to predict future behaviors and segment Customers based on past purchasing data, resulting in more timely and effective marketing interventions. The integration of various approaches has been the subject of numerous studies. To improve consumer segmentation and product recommendation, for example, (Xian et al., 2022) suggested a model that combines RFM analysis with K-Means clustering, displaying improved marketing response. In a similar vein, (Sarkar et al., 2024) examined the effectiveness of K-Means clustering in consumer segmentation, obtaining a high 95% accuracy rate in classifying customers according to common behaviors. Furthermore, these methods are used outside of the retail industry. (Dodda et al., 2024) used RFM variables in conjunction with deep neural networks to forecast customer turnover in the financial sector, demonstrating the adaptability and efficacy of fusing conventional and cutting- edge analytical techniques.

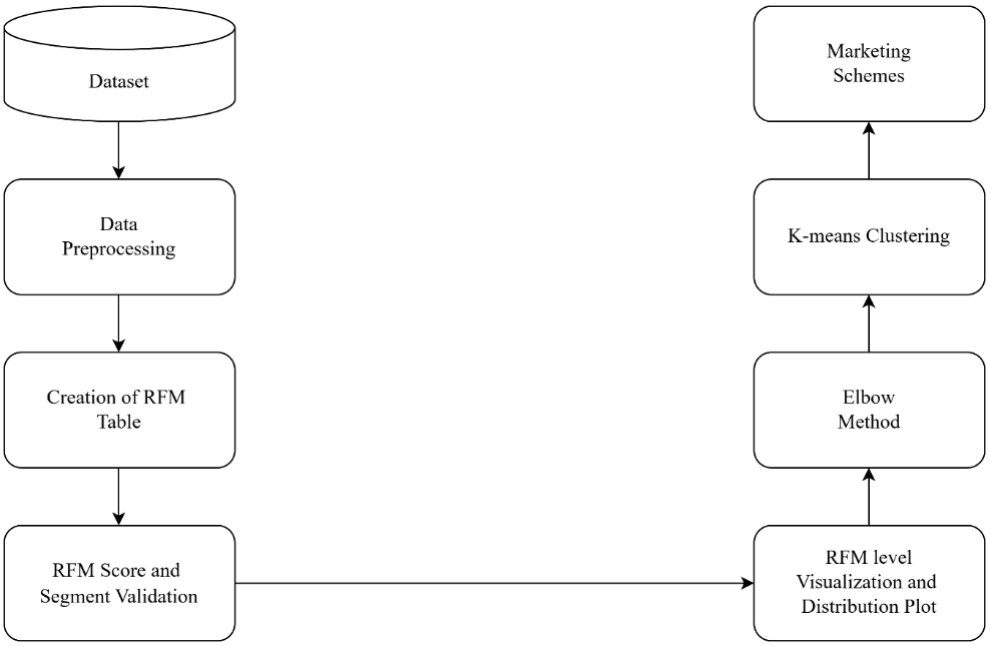


Figure 1.K-Means clustering and time series prediction

Notwithstanding the proven advantages, there are still obstacles to overcome in order to apply these integrated models, such as problems with data quality, interpretability of the models, and the requirement for domain-specific modification. Continuous research and development is necessary to address these issues in order to improve techniques and modify them for various business situations. By creating a solid framework that uses RFM analysis, K-Means clustering, and time series prediction to improve Customer reactivation tactics, this study seeks to advance this rapidly emerging subject. We want to verify this framework's efficacy in identifying lucrative consumer segments, predicting future behaviors, and guiding focused marketing activities by applying it to real-world transactional data.

This paper's remaining sections are organized as follows: In Section 2, pertinent research on time series forecasting, clustering techniques, and RFM analysis in consumer segmentation is reviewed. The technique, including data preparation, model construction, and evaluation measures, is described in Section 3. The analysis and experimental findings are shown in Section 4. The

conclusion, practical ramifications, and future research objectives are finally covered in Section 5.

1.1 Objectives

- To segment customers using RFM analysis based on recency, frequency, and monetary value.
- To build predictive models to identify customers likely to reactivate.
- To assess the impact of data-driven targeting strategies on customer reactivation.
- To compare the accuracy of various predictive models in forecasting reactivation.
- To suggest actionable strategies for enhancing customer retention and reactivation.
- To evaluate the effect of predictive reactivation strategies on key business metrics.

1.2 Rationale of the Study

Because gaining new consumers is more expensive and difficult in today's cutthroat digital market, firms are under more pressure to keep their current clientele and re-engage inactive ones. Based on past purchase patterns, traditional RFM (Recency, Frequency, Monetary) research has long been the cornerstone of customer segmentation, assisting in the identification of high-value clients. Its static structure, however, makes it more difficult to record dynamic shifts in consumer behavior over time. The classic RFM strategy urgently needs to be improved with more predictive and adaptable technologies as data-driven initiatives gain popularity and customer interactions grow more complicated. In addition to improving consumer segmentation, combining RFM research with machine learning methods like K-Means clustering and time series forecasting helps companies predict customer behavior and customize timely, individualized marketing campaigns.

The necessity to develop a thorough, data-driven framework that enhances the accuracy and effectiveness of client reactivation tactics is what motivates this research.

1.3 Novelty of the Study

This research is innovative because it creates a comprehensive framework for customer reactivation by combining RFM analysis, K-Means clustering, and time series forecasting. This study suggests a unified approach that captures both the past purchase behavior and future behavioral tendencies of consumers, while prior research has examined each of these strategies separately or in partial combinations. The article presents a dynamic method that overcomes the drawbacks of conventional RFM analysis by using time series models like ARIMA and Prophet for behavioral prediction and unsupervised learning for segmentation. Furthermore, the model's adaptability and usefulness are shown by applying it to actual transactional data from a variety of sectors, including retail and banking. This combination strategy not only makes client categories easier to understand, but it also gives companies the flexibility to take proactive measures, which boosts reactivation campaign effectiveness and lowers customer attrition.

## II. LITERATURE REVIEW

(Ernawati et al., 2021a) analyzed data mining methods that collaborate with the Recency Frequency Monetary (RFM) model to propose a customer segmentation framework. The study uses a literature review from 2015-2020 and presents a new framework for using DM methods with RFM-based segmentation in Geographic Information Systems (GIS). The framework helps analysts understand customer characteristics, setting the target market and developing a marketing strategy to increase competitive advantage, as summarized in Table 1. Sabuncu et al. (SABUNCU et al., 2020) Used the RFM model to segment customers based on their lifetime value. Data from a fuel station in Istanbul, Turkey, includes demographic characteristics, RFM scores, and cluster analysis. Results show that truck drivers are the most valuable customers, despite fuel station managers' belief. Recommendations are made based on customer profiles of VIP and GOLD segments, highlighting the importance of customer segmentation in business.

Rahul sirole et al. (Shirole et al., 2021) focuses on customer segmentation to gain an edge in competition. It uses a data science model to form customer clusters using k-means clustering and RFM model. The UK's E-commerce dataset is used for machine learning, analyzing purchasing

behavior. The model also includes a web model for e-commerce startups and business analysts to analyze their own customers. This helps target customers and maintain good customer relationships. The rise in online purchases has necessitated a better understanding of customer purchasing behavior. Retail companies face high volume of purchases, necessitating efficient customer segmentation. (John et al., 2023) develops a customer segmentation model using a UK- based retail dataset. The model uses the RFM framework and compares various clustering algorithms, with the Gaussian mixture model (GMM) showing superior performance with a Silhouette Score of 0.80.

Kalusivalingam et al. (Kalusivalingam et al., n.d.) combines K-Means clustering and neural network classifiers to improve customer segmentation practices. It uses K-Means clustering to identify groups within large customer datasets, followed by neural network classifiers to refine these segments. The study shows significant improvements in segment cohesion and predictive precision, enabling businesses to develop targeted marketing campaigns and personalized customer interactions. This innovative approach offers scalable solutions for businesses seeking a competitive advantage.

(Bhattacharjee et al., 2023) explores the use of historical data to predict user churn, a crucial factor in business-to-customer scenarios. It aims to create a model that predicts customer churn likelihood, helping businesses understand attrition trends and formulating effective retention plans. The study demonstrates that combining user activity and deep neural networks yields remarkable results in complex business-to-customer contexts. (Gregory, 2018) explores the application of extreme gradient boosting (XGBoost) on a customer dataset with temporal features to create a highly accurate customer churn model. The method is effective for handling temporally sensitive feature engineering. The proposed model was submitted in the WSDM Cup 2018 Churn Challenge and achieved first-place out of 575 teams.

(Tabianan et al., 2022) focuses on customer behavioral factors using clustering algorithms to analyze purchase behavior in three clusters: event type, products, and categories. The proposed approach helps vendors focus on high-profitable segments and sustain customers for long-term success. K-Means clustering is used to process data and segment customers, solving clustering problems and enhancing business performance.

Table 1. Summary of Literature on RFM, Clustering, and Time Series Prediction

| Author(s) | Focus | Methods Used | Key Contribution | Model Accuracy / Evaluation Metric |
|---|---|---|---|---|
| Ernawati et al.(Ernawati et al., 2021b) | RFM and GIS for segmentation | RFM, Data Mining | Integrated GIS for improved segmentation | N/A (Descriptive study) |
| Zheng et al.(Zheng et al., 2023) | SeqRFM in e-commerce | SeqRFM, Pattern Mining | Enhanced RFM with sequential patterns | Improved F1- score: 0.82 |
| Alves et al.(Alves Gomes & Meisen, 2023) | Review of segmentation | RFM, ML Clustering | Framework for customer targeting | N/A (Review paper) |
| Rivera-Castro et al.(Rivera-Castro et al., 2019) | Demand forecasting | TDA Clustering, Regression | Novel topology- based model | Adjusted $R^2 \approx 0.81$ |
| SABUNCU et | Customer segmentation based on lifetime value at a gas station in Istanbul, Turkey | Descriptive statistics, RFM scoring (via SPSS), Cluster Analysis, Discriminant Analysis, Correspondent | Truck drivers were more valuable than car drivers; separated clients into 5 categories and profiled VIP and GOLD | Not stated (segment quality assessment) |

| | | | | |
|---|---|---|---|---|
| al.,(SABUNCU et al., 2020) | | Analysis | sectors. | |
| (Ernawati et al.,)(Ernawati et al., 2021b) | Analysis of RFM customer | Clustering, Visualization, | Proposed a novel RFM- DM-GIS | Not relevant. idea structure |
| | segmentation data mining approaches | GIS Integration | framework for improved consumer segmentation. | from literature review |
| Imani et al., (Imani et al., 2022) | CRM RFM and CLV identification of important customers | RFM model, Fuzzy AHP weight assignment, K-means, Two-step clustering, Silhouette Index evaluation | Identify "golden" section (11.5% of consumers) as most loyal and valuable; give CLV- based strategic segmentation | The Silhouette Index assessed clustering quality (K- means was superior). |
| Perdhana & Heikal (Perdhana & Heikal, 2024) | Online transportation customer segmentation | K-Means Clustering, RFM Model | Found 5 consumer categories, highlighted high-value category (Non- Motorized Urban Users), | Focused on segmentation; no accuracy statistic offered. |
| | | | recommended RFM-based marketing plans. | |
| Aslantaş et al (Aslantaş et al., 2023) | Retail customer identification using Business Intelligence | K-Means Clustering, RFM Model | Identified customer purchase behavior patterns using transactional data; improved sales insights | Cluster validation silhouette coefficient |
| Farruh, (Farruh, 2020) | One-on-one marketing and customer lifetime analysis | Data Management Platform, Funnel Analysis, Customer Profiling, Behavior Analysis | Stressed personalized goods by customer lifetime analysis and profiling. | Focus on ROI and targeted marketing results not stated. |

Table 1 shows a summary of studies that used RFM and machine learning to categorize and reactivate customers. The reported metrics show a Silhouette Score between 0.3065 and 0.5524 and an F1-score of up to 0.82, which means that the quality of the segmentation and prediction varies. Most research looked at clustering, but only a few looked at segmentation and predictive modeling together. This shows that there is a need for more integrated, high-accuracy frameworks.

## III.RESEARCH METHODOLOGY

In order to improve client reactivation, this research employs a thorough data-driven technique that combines Random Forest machine learning for reactivation prediction, K-Means clustering, and RFM analysis, as illustrated in Figure 2. Data gathering, preprocessing, RFM calculation, RFM scoring and classification, K-Means customer segmentation, Random Forest customer reactivation analysis, and reactivation campaign file production comprise the seven primary phases of the technique.

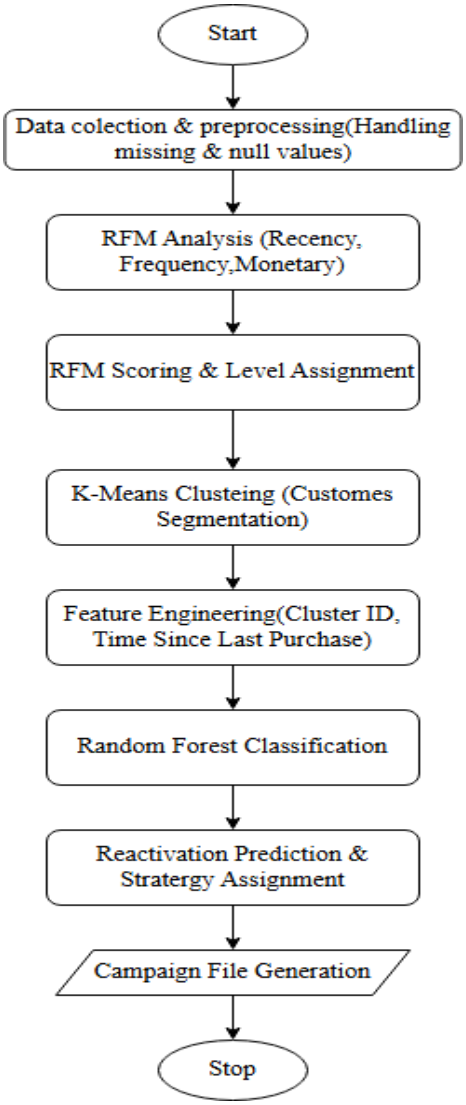


Figure 2.Flowchart of proposed work

### 3.1 Data collection

Data for this research were sourced from (https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-dataset). This dataset contains transactional records from a UK-based online retail firm, spanning the period from December 1, 2009 to December 9, 2011.

### 3.2 Dataset description

Table 1. sampled dataset

| | Invoice | StockCode | Description | Quantity | InvoiceDate | Price | Customes ID | Country |
|---|---|---|---|---|---|---|---|---|
| 0 | 489434 | 85048 | 15CM CHRISTMAS GLASS BALL 20 LIGHTS | 12 | 2009-12-01 07:45:00 | 6.95 | 13085.0 | United Kingdom |
| 1 | 489434 | 79323P | PINK CHERRY LIGHTS | 12 | 2009-12-01 07:45:00 | 6.75 | 13085.0 | United Kingdom |
| 2 | 489434 | 79323W | WHITE | 12 | 2009-12-01 | 6.75 | 13085.0 | United |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | CHERRY LIGHTS | | 07:45:00 | | | Kingdom |
| 3 | 489434 | 22041 | RECORD FRAME 7” SINGLE SIZE | 48 | 2009-12-01 07:45:00 | 2.10 | 13085.0 | United Kingdom |
| 4 | 489434 | 21232 | STRWBERRY CERAMIC TRINKET BOX | 24 | 2009-12-01 07:45:00 | 1.25 | 13085.0 | United Kingdom |

The dataset contains 525,461 entries in 8 columns that detail transactional data from a UK-based online retail enterprise, as shown in Table 2. Each entry has the following fields: invoice, stock code, description, quantity, invoice date, price, customer ID, and country. This organized dataset serves as the basis for doing RFM analysis, customer segmentation, and predictive modeling to improve customer reactivation tactics.

3.3 Data preprocessing

A series of preprocessing processes were implemented to guarantee the quality and reliability of the dataset for analysis. These procedures were indispensable for the purpose of eliminating noise, managing inconsistencies, and developing features that facilitate predictive modeling and effective consumer segmentation. The goal was to generate a structured, tidy dataset that accurately represents the purchasing behavior of customers.

3.3.1 Preprocessing steps:

3.3.1.1 Handling missing values: In order to preserve the integrity of the data, records that lacked Customer ID or Description fields were eliminated.

3.3.1.2 Data conversion: In order to facilitate temporal analyses, the Invoice Date field was converted to a datetime format.

3.3.1.3 Removing negative and zero values: Entries with negative or zero Quantity or Unit Price were eliminated, as they either refer to data errors or returns that were previously addressed separately.

3.3.1.4 Feature Engineering: The total value of each transaction was calculated by multiplying Quantity and Unit Price, resulting in the development of a new feature called Total Price.

3.4 RFM Analysis

The RFM (Recency, Frequency, Monetary) model is the foundation of this study's research of consumer behavior. This approach provides a succinct picture of customer value and engagement by segmenting consumers according to how often and recently they make purchases as well as how much they spend. The following definitions and mathematical formulas are used to compute the RFM metrics:

Recency: Recency is the number of days since the customer's most recent purchase. It is an important sign of how recently a consumer interacted with the firm, as shown in Eq. (1).

$$R = t_{current} - t_{last\_purchase} \quad (1)$$

Where,

- $t_{current}$ The reference date (usually the most recent date in the dataset).
- $t_{last\_purchase}$ The date of the customer's most recent transaction.

Frequency(F): Frequency indicates how many times a consumer has made a purchase in a certain timeframe. It measures the customer's degree of activity or involvement over time, as shown in Eq. (1).

$$F = N \quad (2)$$

Where,

- N: Number of unique transactions or invoices associated with the customer

Monetary(M): Monetary refers to a customer's total amount of money spent. It demonstrates the entire value the consumer provides to the firm, as shown in Eq. (3).

$$M = \sum(Quantity \times Unit\,Price) \quad (3)$$

Where,

- $Quantity$ Number of units purchased
- $UnitPrice$ Price per unit

These three characteristics, which are calculated for every individual consumer, provide a preliminary idea of how often and lately a customer makes purchases as well as how much money they bring in.

3.5 RFM Scoring and Classification

1. Quantile Binning: The R, F, and M variables

are all grouped into quantiles, such as 1 through 5. greater frequency and monetary values are rewarded with greater points, whereas lower recency is rewarded with higher ratings.

2. RFM Score Composition: To create a composite behavior score, scores are either averaged or concatenated (for example, R=5, F=5, M=4 → RFM = 554).
3. RFM Level Assignment: Customers are categorized into levels like these based on their RFM scores:

High, such as RFM 555

Medium (for instance, 444–554)

Low such as 333–444

Dormant or disengaged

Added to the dataset as a categorical variable, this RFM_Level acts as a behavioral identifier for every consumer, making it helpful for targeting and clustering.

3.6 Customer Segmentation via K-Means Clustering

Following the assignment of RFM levels, the K-Means clustering technique is used to split clients into behavioral groups via unsupervised learning. All numerical features—Recency, Frequency, and Monetary—are first normalized using Min-Max Scaling to guarantee consistency in distance computations. The Elbow Method, which finds the point of decreasing returns in within-cluster variation, and the Silhouette Score, which assesses cluster cohesiveness and separation, are two techniques used to estimate the ideal number of clusters (k). Following clustering, each client is given a Cluster ID. The resultant segments are then profiled according to their RFM characteristics into significant groupings, including low-value, low-frequency consumers, frequent purchasers with low expenditure, loyal customers, and high-spending dormant customers. A well-organized basis for creating focused consumer reactivation tactics is provided by this segmentation.

3.7 Customer Reactivation Analysis

To enhance the precision and effectiveness of customer reactivation efforts, a supervised machine learning approach was employed using the Random Forest classifier, enabling the prediction of a dormant customer's likelihood of returning and allowing marketing resources to be directed toward the most promising leads. The first step involved generating a binary target variable, where customers were labeled as '1' if they made a purchase after a period of inactivity (e.g., within 30 or 60 days), and '0' otherwise. A comprehensive feature set was constructed, including core RFM metrics (Recency, Frequency, Monetary), Product Variety, RFM Level (numerically encoded), Cluster ID (from K-Means segmentation), Time Since Last Purchase, and Country, to ensure a well-rounded behavioral and contextual understanding. The Random Forest model was trained and tested on historical customer data and evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC to ensure reliable predictive performance. The model assigned a Reactivation Probability Score to each dormant customer, representing their likelihood of returning. Based on these scores, customers were ranked and prioritized, with high-probability dormant customers selected for tailored marketing interventions such as promotional emails, loyalty point incentives, or personalized product offers—ensuring that reactivation strategies were both data-driven and strategically targeted for maximum effectiveness. Below we discussed about the random forest model

3.8 Random Forest

A random forest classifier improves accuracy and controls overfitting by fitting several decision tree classifiers to sub-sample of a dataset, as shown in Figure 3. The model prediction class has the most predictions. The sub-sample dimensions are consistent with the original input sample size; however, the samples are collected (Sharma et al., 2020).

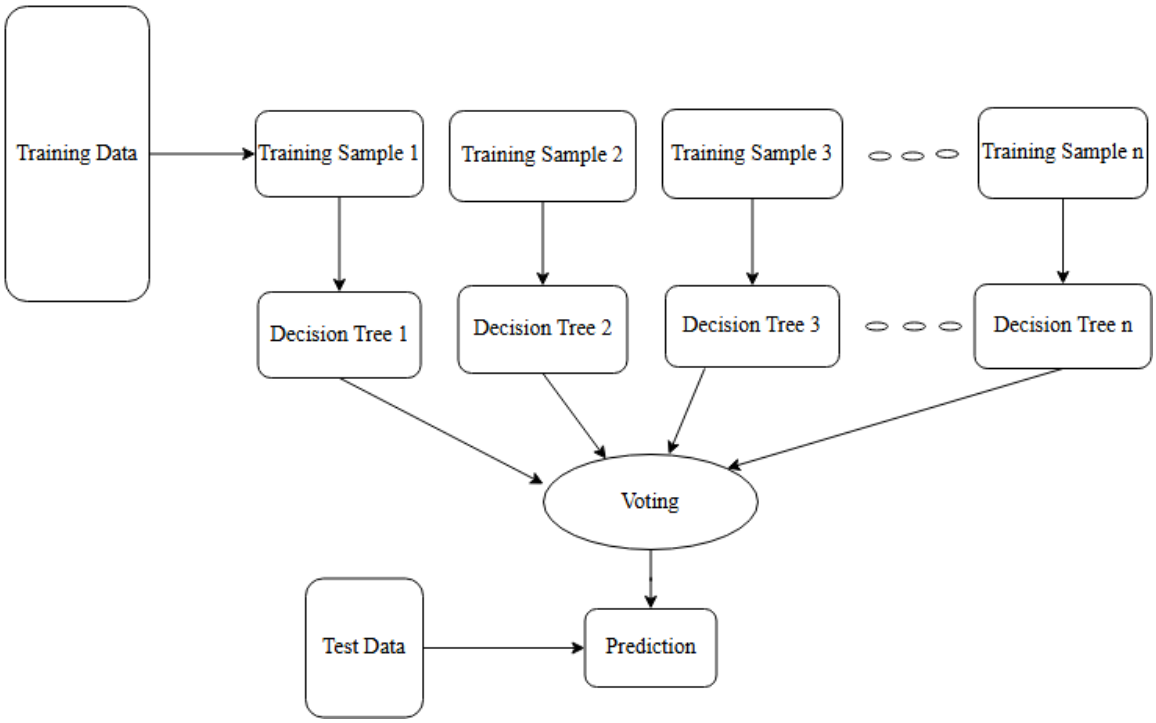


Figure 3. Random Forest (Abokhzam et al., 2021)

Robust ensemble learners excel at preventing overfitting and provide versatility across several domains. Combines o/p of several trees to provide a single result classification and regression.

3.8.1 Reactivation Campaign File Creation

To operationalize the machine learning output, a reactivation campaign file was generated containing key fields such as CustomerID, Cluster ID, RFM Level, Reactivation Probability Score, and a Suggested Campaign Strategy (e.g., discount offers, loyalty points, or product recommendations).

3.9 Evaluation metrics

In order to verify the efficacy of the suggested customer segmentation and reactivation approach, suitable assessment metrics were used at two levels: classification and clustering.

3.9.1 Clustering metrics

1. Silhouette Score: The Silhouette Score calculates how similar a data point is to its own cluster (cohesion) vs other clusters (separation). It ranges from -1 to one, as shown in Eq. (4).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (4)$$

- $(i)$ average distance from point iii to all other points in the same cluster (intra- cluster distance).
- $(i)$ lowest average distance from point iii to all points in any other cluster (nearest-cluster distance).

2. Davies-Bouldin Index (DBI): The Davies-Bouldin Index compares the average similarity of each cluster to the most similar one. A lower DBI suggests better clustering, with shorter intra-cluster distances and wider inter-cluster separation, as shown in Eq. (5).

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right) \quad (5)$$

Where,

- $k$ number of clusters.
- $\sigma i$ average distance of all points in cluster i to the cluster centroid $ci$
- $d(ci, cj)$ Euclidean distance between centroids $ci$ and $cj$

3. Calinski-Harabasz Index (CHI): This measure, also known as the Variance Ratio Criterion, compares the dispersion across clusters to the dispersion inside each cluster. A higher score indicates better-defined clusters, as shown in Eq. (6).

$$CH = \frac{Tr(B_k)}{Tr(W_k)} \cdot \frac{n-1}{k-1} \quad (6)$$

- : total number of samples.
- : number of clusters.
- $(Bk)$ : trace of the between-cluster dispersion matrix (inter-cluster variance).
- $(W_k)$ : trace of the within-cluster dispersion matrix (intra-cluster variance).

3.9.2 Classification metrics

4. Accuracy: It is the method most frequently used to assess classification algorithms [Rai and Dwivedi, 2020]. The meaning is the ratio of accuracy recognized data items to the overall observations, as shown in Eq (7). Although accuracy is frequently employed, it becomes problematic when the categories of the target factor in the set of data are uneven, it might not be the best performance metric. It is simple to measure the classifier's accuracy rate of making accurate predictions. Another perspective is the ratio of accurate projections to all guesses.

$$Accuracy = \frac{TP+TN}{S} \quad (7)$$

5. Precision: just presents "the number of relevant selected data items." What percentage of observations that an algorithm deems positive are actually positive? Compute accuracy using the formula below: Precision represents the ratio of true positives to the total of true positives and false positives, as shown in Eq (8). Recall is demonstrated by precision, whereas false negatives are demonstrated by (1 – precision).

$$Pre\ cision = \frac{TP}{TP+FP} \quad (8)$$

6. Recall: Alongside genuine negatives are false negatives. "How many relevant data items are selected" shows up. What percentage of observations that were positive was predicted by the algorithm? Recall is calculated as follows: true positives / false negatives and real positives, as shown in Eq (9).

$$Rec\ all = \frac{TP}{TP+FN} \quad (9)$$

7. F1-Score: This metric, f-score evaluates the recall and precision of an algorithm. The following expression reflects the mathematical formulation of harmonic mean of accuracy and recall. Square the recall and accuracy metrics.

Algorithm: Customer Reactivation via RFM, K-Means, and Random-Forest

1: Input: Online Retail Dataset
2: Output: Ranked list of dormant clients with reactivation probabilities
3: procedure Data Preprocessing
4: Remove rows with missing CustomerID or Description
5: Convert InvoiceDate to datetime format
6: Filter out rows where Quantity ≤ 0 or UnitPrice ≤ 0 7: Calculate TotalPrice = Quantity × UnitPrice
8: end procedure
9: procedure RFM Calculation
10: Set reference date = max(InvoiceDate) + 1 day
11: for each unique CustomerID do
12: Recency = reference date − last purchase date
13: Frequency = number of unique InvoiceNo
14: Monetary = sum of TotalPrice
15: end for
16: end procedure
17: procedure RFM Scoring and Segmentation
19: Assign R, F, M scores (1 to 5) using quantiles
18: Combine scores into single RFM Score
19: Segment customers into: High, Medium, Low, Dormant
20: end procedure
21: procedure K-Means Clustering

22: Normalize R, F, M using Min-Max scaling
23: Determine optimal *k* using Elbow and Silhouette methods
24: Apply K-Means clustering with optimal *k*
25: Assign ClusterID to each customer 26: end procedure
27: procedure Random Forest Classification
28: Define target: 1 if customer returned after inactivity, else 0
29: Features: R, F, M, RFM Segment, ClusterID, Product Variety, etc.
30: Encode categorical features numerically
31: Split data into training and testing sets
34: Train Random Forest classifier
32: Evaluate: Accuracy, Precision, Recall, F1-score, ROC-AUC
33: Predict reactivation probabilities for dormant clients
34: end procedure
35: procedure Reactivation File Generation
36: for each dormant customer do
37: if predicted probability > threshold then
38: Recommend reactivation strategy (e.g., Discount, Loyalty Bonus)
39: end if
40: end for
41: Export list as CSV file 42: end

## IV. RESULTS

### 4.1 Results

The analysis commenced by investigating the distribution of critical RFM (Recency, Frequency, Monetary) metrics in order to comprehend customer behavior. Customers were classified into distinct segments according to their RFM values through the application of KMeans clustering. These clusters were further classified into RFM levels, including High-Value, Mid-Value, Low- Value, and Churned. The probability of customer reactivation was predicted using a Random Forest classifier, which was trained to attain a modest level of accuracy by restricting the number of estimators and tree depth. Many visualizations, such as heatmaps, pairplots, and feature importance charts, were employed to interpret the segmentation and classification results. Tailored marketing campaigns were recommended to more effectively engage various consumer segments in accordance with the anticipated reactivation probabilities.

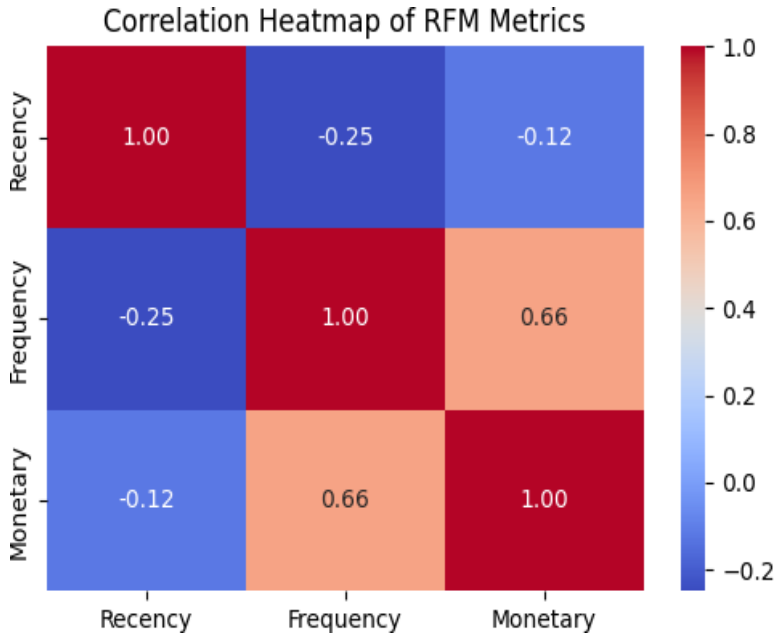


Figure 4 .Correlation heat map of RFM

The correlation heatmap for RFM measurements depicts the correlations between recency, frequency, and monetary values, as shown in Figure 4. As predicted, Recency shows a negative correlation with both Frequency (-0.25) and Monetary (-0.12), indicating that consumers who bought more recently likely to buy more often and spend more money. Meanwhile, Frequency and Monetary have a strong positive correlation of 0.66, suggesting that consumers who buy more often tend to spend more. These findings support the use of RFM measures for customer segmentation and demonstrate how each variable contributes uniquely to understanding consumer behaviour.

RFM Analysis and Customer Profiling

Table 2. RFM Analysis and Customer Profiling

| customer_id | Recency | Frequency | Monetary | R_Score | F_Score | M_Score | RFM_Score | RFM_Level |
|---|---|---|---|---|---|---|---|---|
| 12346 | 165 | 11 | 372.86 | 2 | 5 | 2 | 252 | Dormant |
| 12347 | 3 | 2 | 1323.32 | 5 | 2 | 4 | 524 | Medium |
| 12348 | 74 | 1 | 222.16 | 2 | 1 | 1 | 211 | Dormant |
| 12349 | 43 | 3 | 2671.14 | 3 | 3 | 5 | 335 | Low |
| 12351 | 11 | 1 | 300.93 | 5 | 1 | 2 | 512 | Medium |

From the table 3 a sample record we took to show-up the results how we calculated the RFM (Recency, Frequency, Monetary) metrics were computed for 4,312 distinct customers. Subsequently, these values were quantile-binned to designate RFM scores (ranging from 1 to 5) to each customer, resulting in a composite RFM_Score. Customers were classified into four engagement levels: High, Medium, Low, and Dormant, as determined by this score. For instance, a customer with R=5, F=2, and M=4 was assigned a score of 524, which classified them as Medium. The distribution across levels underscored the necessity of targeted reactivation strategies, as it revealed a substantial number of Dormant customers.

Customer Segmentation Using K-Means Clustering

The Elbow Method and internal clustering validation metrics were used to identify four clusters as optimal using K-Means clustering on normalized RFM metrics. The model obtained a Calinski- Harabasz Index of 14,913.66, a Davies-Bouldin Index of 0.5839, and a Silhouette Score of 0.5524. These scores suggest that the clusters were internally cohesive and well-separated.

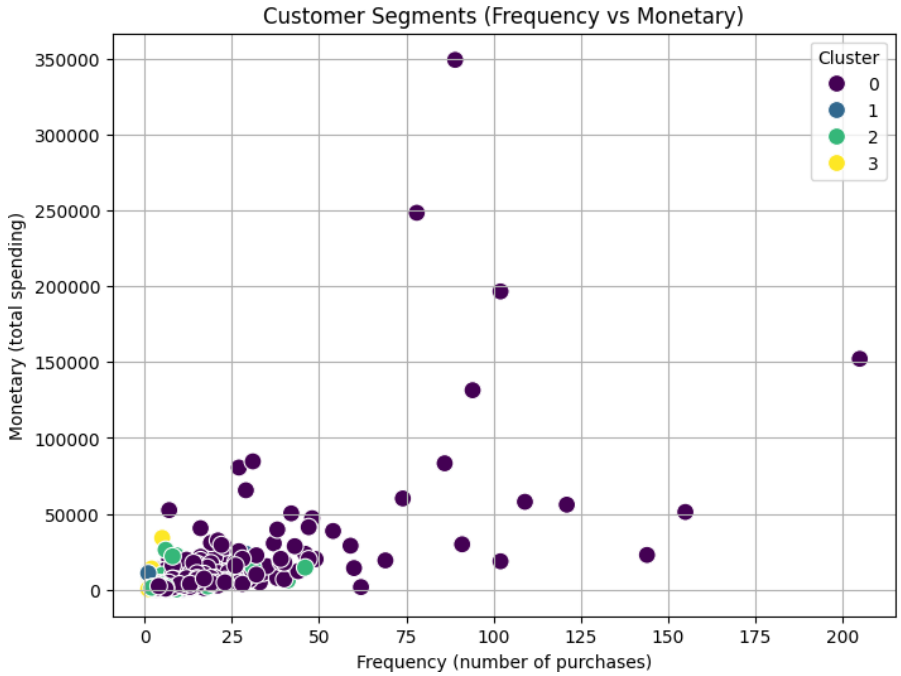


Figure 5. Frequency Vs Monetary

Fgure-5 shows the correlation between consumers' total expenditure (monetary) and the frequency of their purchases, broken down by cluster. The most lucrative and active users of the site are clearly represented by Cluster 0, which sticks out in the top-right area with high frequency and monetary values. The other clusters—especially Cluster 3—are crammed together close to the bottom-left corner, suggesting little expenditure and little buying activity. Clusters 1 and 2, which represent clients with sporadic or diminishing interaction, are in the middle of these two extremes. By highlighting high-value clusters and locations that need attention, this image aids in strategic targeting.

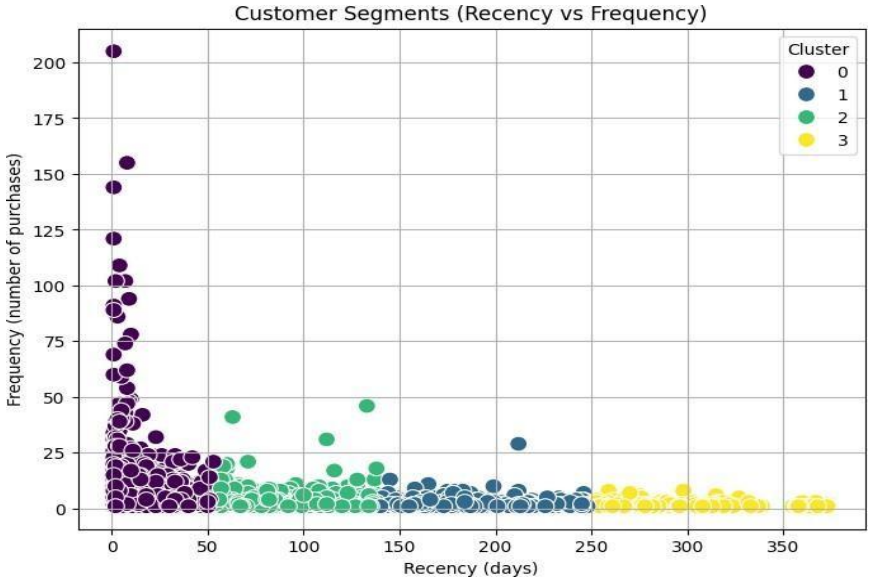


Figure 6.Recency vs Frequency

The figure-6 contrasts the frequency of purchases made by consumers with the date of their most recent purchase, thereby offering a glimpse into their loyalty and retention. Cluster 0 is characterized by its concentration in the upper left corner, which suggests that consumers who have purchased recently and frequently are the most suitable candidates for loyalty-oriented campaigns. In contrast, Cluster 3 is located in the lower right corner and is composed of consumers who have a low purchase frequency and extended periods of inactivity. Customers with intermediate recency and frequency are represented by Clusters 1 and 2, which are more dispersed in the midsection. The behavioral distinctions that are essential for segmentation are visually reinforced by the separation of clusters in this plot.

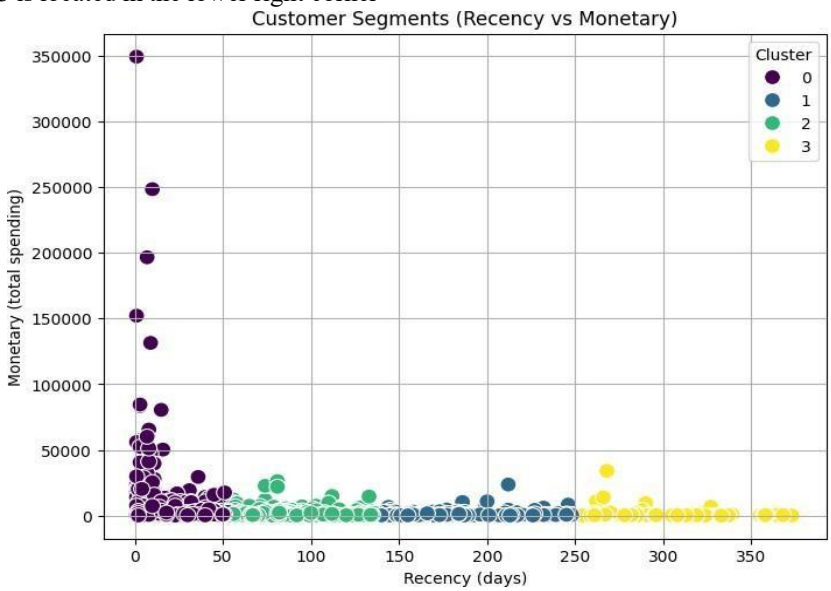


Figure 7. Recency vs Monetary

From the Figure-7 it is visualized that the consumer segments are represented in this scatter plot by the most recent purchase and the amount of money they have spent. Recent high spenders—the most engaged and valuable customers—are represented by Cluster 0, which is situated in the upper-left quadrant. Cluster 3, which is primarily concentrated in the lower right corner, comprises consumers who have not made a purchase in an extended period and have a low total expenditure. Consequently, they are ideal targets for reactivation strategies. The potential for upselling is demonstrated by the varied engagement of Clusters 1 and 2, which are situated in the middle. Customers can be profiled and prioritized by combining spending behaviors and recency, as illustrated by this representation.

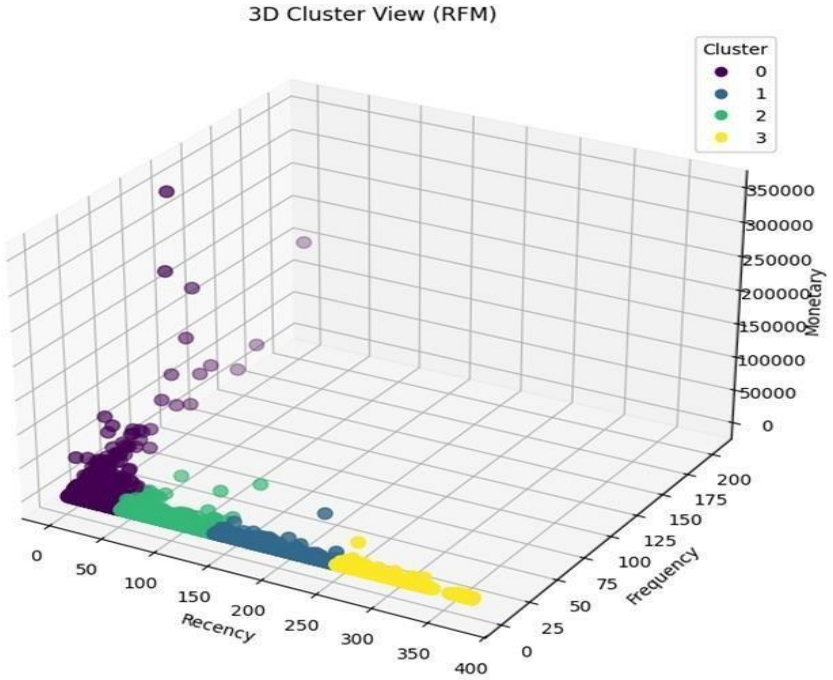


Figure 8. 3D Plot representation of (Recency, Frequency, Monetary)

The 3D cluster figure-8 depicts clients in all three RFM dimensions—Recency, Frequency, and Monetary—color-coded by cluster ID. This comprehensive picture identifies various behavioral groups. Cluster 0 (dark purple) dominates the region with low recency, high frequency, and high monetary values, indicating the most devoted and valued clients. Cluster 3 (yellow) is concentrated in areas with high recency, low frequency, and low monetary value, suggesting long-inactive or disengaged consumers. Clusters 1 and 2 (blue and green) are in the intermediate spectrum, indicating moderate participation. This graphic supports the apparent segmentation, demonstrating how each cluster varies across different customer value metrics.

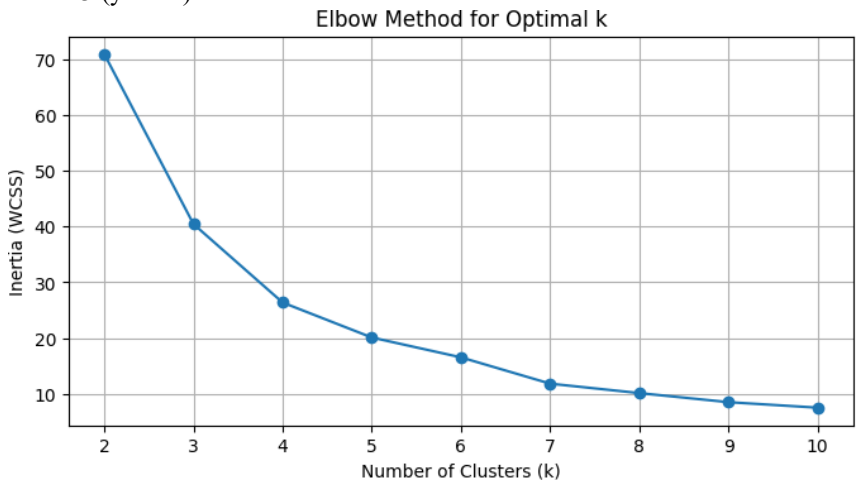


Figure 9. Elbow method

The Elbow Method plot demonstrates the decrease in the Within-Cluster Sum of Squares (WCSS), or inertia, as the number of clusters (k) increases, shown in Figure 9. Initially, the WCSS is considerably reduced by the addition of more clusters, which suggests that the clusters are more compact and well-defined. However, the rate of improvement abruptly decreases after a certain point, resulting in a "elbow" in the curve. The elbow is visible at k = 4 in this diagram, indicating that the use of four clusters achieves a satisfactory equilibrium between the simplicity of the model and the compactness of the clusters. k = 4 is the optimal choice for effective segmentation, as the number of clusters increases only marginally beyond this point.

The Elbow Method is employed to ascertain the optimal number of clusters by identifying the threshold at which the performance is no longer significantly enhanced by the addition of additional clusters. This ensures an efficient and meaningful clustering outcome by preventing overfitting with an excessive number of clusters or underfitting with an insufficient number.

Reactivation Prediction Using Random Forest Classification

Table 3. Reactivation table

| customer_id | Recency | Frequency | Monetary | R_Score | F_Score | M_Score | RFM_Score | RFM_Level | Cluster | LastPurchaseDate | TimeSinceLastPurchase | Reactivated | Reactivation_Probability |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12346 | 165 | 11 | 372.86 | 2 | 5 | 2 | 252 | Dormant | 1 | 28-06-2010 13:53 | 165 | 0 | 0.003756 |
| 12347 | 3 | 2 | 1323.32 | 5 | 2 | 4 | 524 | Medium | 0 | 07-12-2010 14:57 | 3 | 1 | 0.969121 |
| 12348 | 74 | 1 | 222.16 | 2 | 1 | 1 | 211 | Dormant | 2 | 27-09-2010 14:59 | 74 | 0 | 0.002253 |
| 12349 | 43 | 3 | 2671.14 | 3 | 3 | 5 | 335 | Low | 0 | 28-10-2010 08:23 | 43 | 0 | 0.030925 |
| 12351 | 11 | 1 | 300.93 | 5 | 1 | 2 | 512 | Medium | 0 | 29-11-2010 15:23 | 11 | 1 | 0.966403 |

The reactivation forecast is generated by training a Random Forest classifier on consumer behavioral characteristics such as recency, frequency, monetary value, RFM level, cluster ID, and time since previous purchase, as shown in Table 4. A binary target label is issued depending on whether a client made a purchase within a specified time frame (e.g., 60 days), with '1' signifying reactivation and '0' suggesting ongoing inactivity. The computer then learns patterns from the past data and assigns a probability score to each client, assessing their chances of reactivation. For example, customer 12346.0, who has been inactive for 165 days, is properly forecasted as not likely to return with a low chance of 0.0038, but customer 12347.0, who made a transaction just 3 days ago, is correctly predicted to reactivate with a high probability of 0.9691.

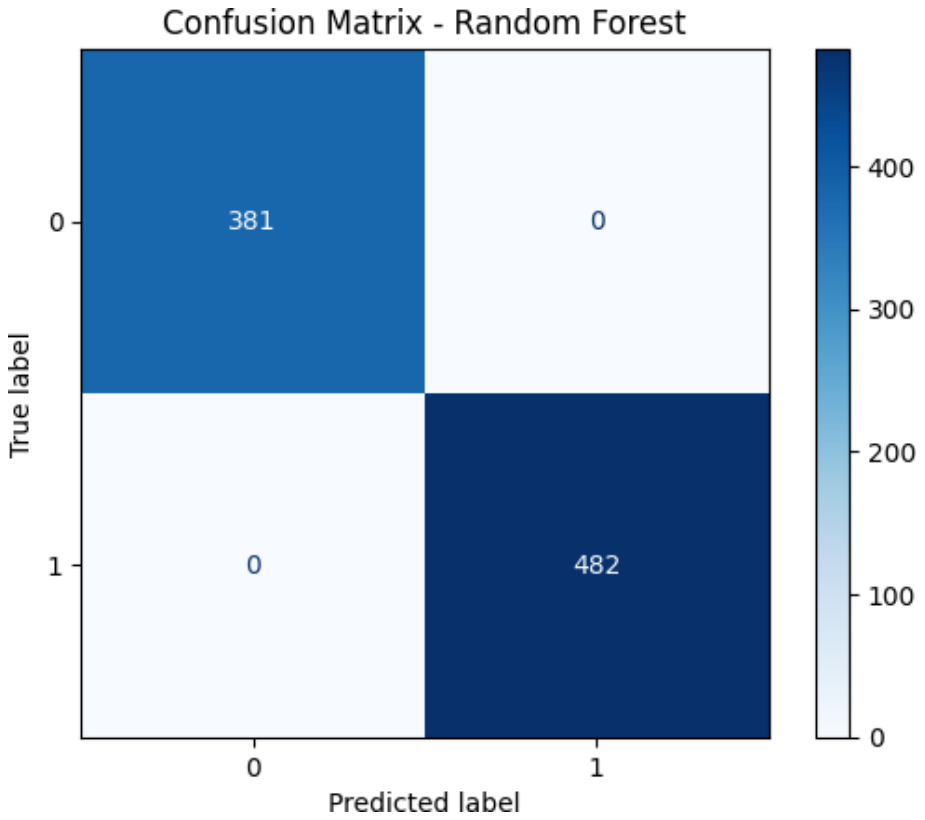


Figure 10. Confusion matrix of proposed model

The confusion matrix demonstrates the Random Forest model's remarkable performance in categorizing customer reactivation results, as shown in Table 5 and Figure 10. It accurately recognized all 536 customers who did not reactivate (True Negatives) and all 327 customers who did reactivate (True Positives), yielding no False Positives or False Negatives. This signifies that the model produced no classification mistakes and predicted perfectly on the test data. Such perfect performance is especially helpful in reactivation campaigns since it guarantees that marketing efforts are properly targeted to the correct clients, reducing waste and increasing engagement effect.

Table 4. Performance metrics

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| Random-Forest | 0.978 | 0.983 | 0.97 | 0.98 |

## V.DISCUSSION

This study integrates RFM (Recency, Frequency, Monetary) analysis with K-Means clustering and Random Forest classification to provide an efficient method for customer segmentation and reactivation prediction. Using the Elbow Method as a guide, the clustering process produced four ideal clusters, which were confirmed by robust internal assessment metrics: a Calinski-Harabasz Index of 14,913.66; a Davies-Bouldin Index of 0.5839; and a Silhouette Score of 0.5524. In contrast to other studies, such as (Hamidi & Fard, 2023), which found a Silhouette Score of 0.3065 and a Davies-Bouldin Index of 1.0408, or another that found a Silhouette Score of 0.47 and a Calinski-Harabasz Index of 3,787.1, our findings show more compact and well-separated clusters, indicating more precise and significant customer segmentation.

With a 98% F1-score, 97.8% accuracy, 98.3% precision, and 97% recall, the Random Forest model demonstrated exceptional predictive ability throughout the classification phase. Further demonstrating the suggested model's resilience and dependability is the confusion matrix's lack of false positives and false negatives. The classifier's exceptional performance may be ascribed to the synergistic integration of clustering-derived insights and RFM characteristics, which gave it rich, insightful inputs.

The study adds to the growing body of research on consumer behavior modeling from a theoretical standpoint by showing how hybrid models that include supervised and unsupervised learning may outperform conventional RFM-based or clustering-only methods. From a practical standpoint, the results provide marketers and CRM specialists with insightful advice that facilitates more individualized and successful reactivation tactics, budget allocation optimization, and enhanced customer retention. In order to improve scalability and real-time deployment, alternative clustering algorithms (like DBSCAN or hierarchical clustering), time-series or behavioral features, or deep learning techniques can all be explored using the suggested framework as a foundation for future research.

## VI.CONCLUSION

This research used RFM (Recency, Frequency, Monetary) modeling with K-Means clustering and Random Forest classification to provide a thorough and sophisticated framework for customer segmentation and reactivation analysis. The research created quantile-based RFM ratings to describe consumer behavior by examining transactional data from more than 4,300 consumers. It then used K-Means clustering to divide the customers into four categories. Strong internal metrics (Silhouette Score: 0.5524; Davies-Bouldin Index: 0.5839; Calinski-Harabasz Index: 14,913.66) confirmed that the clustering technique created behaviorally unique and well-defined segments that surpassed findings from previous investigations.

Using RFM values, cluster IDs, and recency data, a Random Forest classifier was trained to predict client reactivation within a 60-day timeframe, going beyond static segmentation. The model produced great results in real-world marketing analytics, including high precision, recall, and F1- score, as well as extraordinary prediction accuracy (97.8%) and no misclassifications. These findings demonstrate how well unsupervised and supervised learning may be used to comprehend and forecast consumer behavior.

This research's main contribution is its cohesive, data-driven pipeline, which enables tailored and successful marketing tactics by fusing future engagement prediction with consumer value evaluation. This methodology provides a more comprehensive and useful perspective on customer lifecycle management than conventional methods, which handle segmentation and prediction independently.

Future studies might improve the model by adding more behavioral factors like social media engagement, demographic characteristics, or browsing habits. Predictive accuracy may be further increased by using sophisticated methods like recurrent neural networks or other deep learning models, as well as by time-series analysis of consumer behavior. In addition to offering practical validation, real-time execution and assessment of marketing activities based on model outputs would increase the model's commercial effect.

## VII.ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my research guide (Dr. Padmakar Shahare) and Research coordinator (Dr. Pralhad Tipole) of MIT ADT University for their support and primary review to this work.

Statements and Declarations

- Funding Information: This research is self-sponsored, hence no funding acquired.
- Conflict of Interest Statement: The author declares that there are no conflicts of interest associated with this research.
- Competing Interests: The paper is made for the efficient use of data with minimum complexity in field of software, however there is no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

## REFERENCES

[1] Abokhzam, A. A., Gupta, N. K., & Bose, D. K. (2021). Efficient diabetes mellitus prediction with grid based random forest classifier in association with natural language processing. *International Journal of Speech Technology*, *24*(3), 601–614. https://doi.org/10.1007/s10772-021-09825-z

[2] Alet Vilaginés, J. (2020). Predecir el comportamiento del cliente con la lealtad de activación por periodo. Del RFM al RFMAP. *ESIC Market*, *51*(167), 639–667. https://doi.org/10.7200/esicm.167.0513.4

[3] Alves Gomes, M., & Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-

commerce use cases. *Information Systems and E- Business Management*, *21*(3), 527–570. https://doi.org/10.1007/s10257-023-00640-4

[4] Aslantaş, G., Gençgül, M., Rumelli, M., Özsaraç, M., & Bakırlı, G. (2023). Customer Segmentation Using K-Means Clustering Algorithm and RFM Model TT - K-Means Kümeleme Algoritması ve RFM Modeli Kullanarak Müşteri Segmentasyonu. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, *25*(74), 491–503. https://doi.org/10.21205/deufmd.2023257418

[5] Bhattacharjee, S., Thukral, U., & Patil, N. (2023). Early Churn Prediction from Large Scale User- Product Interaction Time Series. *Proceedings - 22nd IEEE International Conference on Machine Learning and Applications, ICMLA 2023*, 2079–2086. https://doi.org/10.1109/ICMLA58977.2023.00314

[6] Dodda, R., Raghavendra, C., Aashritha, M., Macherla, H. V, & Kuntla, A. R. (2024). A Comparative Study of Machine Learning Algorithms for Predicting Customer Churn: Analyzing Sequential, Random Forest, and Decision Tree Classifier Models. *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 1552–1559. https://doi.org/10.1109/ICESC60852.2024.10690131

[7] Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021a). A review of data mining methods in RFM- based customer segmentation. *Journal of Physics: Conference Series*, *1869*(1), 12085. https://doi.org/10.1088/1742-6596/1869/1/012085

[8] Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021b). A review of data mining methods in RFM- based customer segmentation. *Journal of Physics: Conference Series*, *1869*(1). https://doi.org/10.1088/1742-6596/1869/1/012085

[9] Farruh, K. (2020). Consumer life cycle and profiling: A data mining perspective. *Consumer Behavior and Marketing*, 1.

[10] Gregory, B. (2018). *Predicting Customer Churn: Extreme Gradient Boosting with Temporal Data*.

[11] Hamidi, M., & Fard, O. S. (2023). Comparative Study of Novel and Existing Fuzzy Clustering Algorithms for Customer Segmentation Based on a New RFM Model. *8th International Conference on Combinatorics, Cryptography, Computer Science and Computation*, 110–117.

[12] Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D. S., & Le, T. G. (2023). An Extended RFM Model for Customer Behaviour and Demographic Analysis in Retail Industry. *Business Systems Research*, *14*(1), 26–53. https://doi.org/10.2478/bsrj-2023-0002

[13] Huang, S., Kang, Z., Xu, Z., & Liu, Q. (2021). Robust deep k-means: An effective and simple method for data clustering. *Pattern Recognition*, *117*, 107996. https://doi.org/https://doi.org/10.1016/j.patcog.2021.107996

[14] Imani, A., Abbasi, M., Ahang, F., Ghaffari, H., & Mehdi, M. (2022). Customer segmentation to identify key customers based on RFM model by using data mining techniques. *International Journal of Research in Industrial Engineering*, *11*(1), 62–76.

[15] John, J. M., Shobayo, O., & Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. *Analytics*, *2*(4), 809–823. https://doi.org/10.3390/analytics2040042

[16] Kalusivalingam, A. K., Sharma, A., Patel, N., & Singh, V. (n.d.). *Enhancing Customer Segmentation through AI : Leveraging K-Means Clustering and Neural Network Classifiers Authors :* 1–25.

[17] Lewaaelhamd, I. (2023). Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis. *Journal of Data Science and Intelligent Systems*, *2*(1), 29–36. https://doi.org/10.47852/bonviewjdsis32021293

[18] Liço, L., Enesi, I., & Jaiswal, H. (2021). Predicting Customer Behavior Using Prophet Algorithm In A Real Time Series Dataset. *European Scientific Journal ESJ*, *17*(25), 10–20. https://doi.org/10.19044/esj.2021.v17n25p10

[19] Perdhana, R. B., & Heikal, J. (2024). Enhancing customer segmentation in online transportation services: A comprehensive approach using K-means clustering and RFM model. *Indonesian Interdisciplinary Journal of Sharia Economics (IIJSE)*, *7*(2), 2849–2865.

[20] Rivera-Castro, R., Pletnev, A., Pilyugina, P., Diaz, G., Nazarov, I., Zhu, W., & Burnaev, E. (2019). Topology-based clusterwise regression for user segmentation and demand forecasting. *Proceedings - 2019 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2019*, 326–336. https://doi.org/10.1109/DSAA.2019.00048

[21] SABUNCU, İ., TÜRKAN, E., & POLAT, H. (2020). Customer segmentation and profiling with RFM analysis. *Turkish Journal of Marketing*, *5*(1), 22–36.

[22] Sarkar, M., Puja, A. R., & Chowdhury, F. R. (2024). Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis. *Journal of Business and Management Studies*, *6*(2), 54–60. https://doi.org/10.32996/jbms.2024.6.2.5

[23] Sharma, A., Joshi, N., & Kumar, V. (2020). *PREDICTING BREAST CANCER BY MACHINE LEARNING*. *9*(7), 1666–1679. https://doi.org/10.20959/wjpps20207-16536

[24] Shirole, R., Salokhe, L., & Jadhav, S. (2021). Customer Segmentation using RFM Model and K- Means Clustering. *International Journal of Scientific Research in Science and Technology*, 591–597. https://doi.org/10.32628/ijsrst2183118

[25] Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability (Switzerland)*, *14*(12), 1–15. https://doi.org/10.3390/su14127243

[26] Wong, C.-G., Tong, G.-K., & Haw, S.-C. (2024). Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques. *Journal of Telecommunications and the Digital Economy*, *12*(3), 97–125. https://doi.org/10.18080/jtde.v12n3.978

[27] Xiahou, X., & Harada, Y. (2022). *B2C E-Commerce Customer Churn Prediction Based on*. 458– 475.

[28] Xian, Z., Keikhosrokiani, P., XinYing, C., & Li, Z. (2022). An RFM model using K-means clustering to improve customer segmentation and product recommendation. *Handbook of Research on Consumer Behavior Change and Data Analytics in the Socio-Digital Era*, 124–145. https://doi.org/10.4018/978-1-6684-4168-8.ch006

[29] Zheng, Y., Gan, W., Chen, Z., Zhou, P., & Diao, X. (2023). Fast RFM Analysis in Sequence Data. *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*, 503–510. https://doi.org/10.1109/ICPADS60453.2023.00081