

Optimizing Marketing Campaigns Using Random Forest and A/B Testing

Murugeshwari P.¹, Yuvaraj R.^{2,*}, Yogeshwaran M.², Prasanna Kumar U.²

Abstract

Marketing initiatives play a vital role in driving business growth by reaching targeted consumer segments through tailored strategies across multiple channels. The success of these initiatives is influenced by various factors, including the type and duration of the campaign, the characteristics of the target audience, the communication channels employed, and the overall efficiency of each strategy. These factors collectively impact key performance metrics such as conversion rates, customer acquisition costs, and return on investment (ROI). This study explores a comprehensive dataset containing parameters such as Campaign ID, Company Name, Campaign Type, Target Audience, Campaign Duration, Channel Used, Conversion Rate, Acquisition Cost, ROI, and Location. By applying machine learning techniques, specifically the Random Forest algorithm, this research aims to uncover hidden patterns and actionable insights from the data. The primary objectives include identifying the most cost-effective marketing channels, understanding how audience segmentation influences campaign outcomes, and determining the optimal configurations for successful campaigns. The Random Forest model is particularly suited for capturing complex, nonlinear relationships in the data and producing reliable predictions. These insights can be used to refine marketing strategies, enhance decision-making processes, allocate resources more efficiently, and ultimately maximize the overall impact and profitability of marketing campaigns.

Keywords: Marketing campaign optimization, Random forest algorithm, audience segmentation, ROI prediction, data driven insights, A/B testing

INTRODUCTION

A key component of contemporary company strategy are marketing campaigns, which are intended to advertise goods, services, or concepts to certain target markets. They include a wide range of strategies, including traditional media like print and television as well as digital ads, email marketing,

and social media promotions. For businesses to accomplish their goals, which include raising brand recognition, boosting sales, and optimizing return on investment (ROI), these campaigns must be effective. A major obstacle, nevertheless, is the intricacy of examining several variables, including campaign style, target audience, duration, and channel performance. For managing such complexity, machine learning algorithms, Random Forest in particular, have become extremely effective tools.

These algorithms may find trends, forecast results, and offer useful insights by examining big datasets. The goal of this study is to optimize budget allocation, pinpoint important success

*Author for Correspondence

Yuvaraj R.

E-mail: yuvi.2003kce@gmail.com

¹Professor, Artificial Intelligence and Data Science, Karpagam College of Engineering (Anna University) Coimbatore, Tamil Nadu, India

²Student, Artificial Intelligence and Data Science, Karpagam College of Engineering (Anna University) Coimbatore, Tamil Nadu, India

Received Date: March 24, 2025

Accepted Date: May 12, 2025

Published Date: August 07, 2025

Citation: Murugeshwari P., Yuvaraj R., Yogeshwaran M., Prasanna Kumar U. Optimizing Marketing Campaigns Using Random Forest and A/B Testing. Journal of Artificial Intelligence Research & Advances. 2025; 12(3): 1–9p.

elements, and enhance overall marketing effectiveness by analyzing campaign data using the Random Forest algorithm.

Marketing Campaign Optimization

In order to get the greatest results, marketing campaign optimization entails carefully examining and modifying several campaign elements. This approach considers a number of important elements, such as the campaign's kind (e.g., email, social media, television), target audience, duration, channel choice, and budget. In order to optimize key performance indicators (KPIs) including conversion rates, client acquisition expenses, and total return on investment (ROI), these components must be refined. By evaluating historical campaign data and forecasting which tactics are most likely to produce favorable results, machine learning models such as Random Forest can improve this procedure. These models may be used to determine the best channels, target audiences more precisely, and recommend changes to the timing and content of campaigns. Businesses may cut down on unnecessary expenditure and gradually enhance campaign effectiveness through ongoing optimization, which will eventually increase marketing spend efficiency and provide better outcomes. Continuous A/B testing and real-time data analysis further refine campaign strategies, ensuring that marketing efforts stay highly relevant, adaptive, and consistently effective.

Random Forest Algorithm

A potent ensemble learning technique, the Random Forest algorithm constructs many decision trees and aggregates their results to provide a final forecast. By combining predictions from several decision trees, Random Forest reduces the danger of overfitting to particular data points, improving accuracy and generalization. The approach is especially well-suited for managing big and complicated datasets since each tree in the forest is constructed using a random subset of the data and characteristics. It is adaptable for jobs like forecasting campaign results like conversion rates, audience engagement, and return on investment (ROI) since it can handle both regression and classification issues. Random Forest works well in marketing because it may reveal complex correlations between factors, such how different campaign kinds and audience groups interact. Businesses may focus their efforts on the areas that have the most impact by using insights on the relative value of various aspects, which promotes optimization across marketing campaigns.

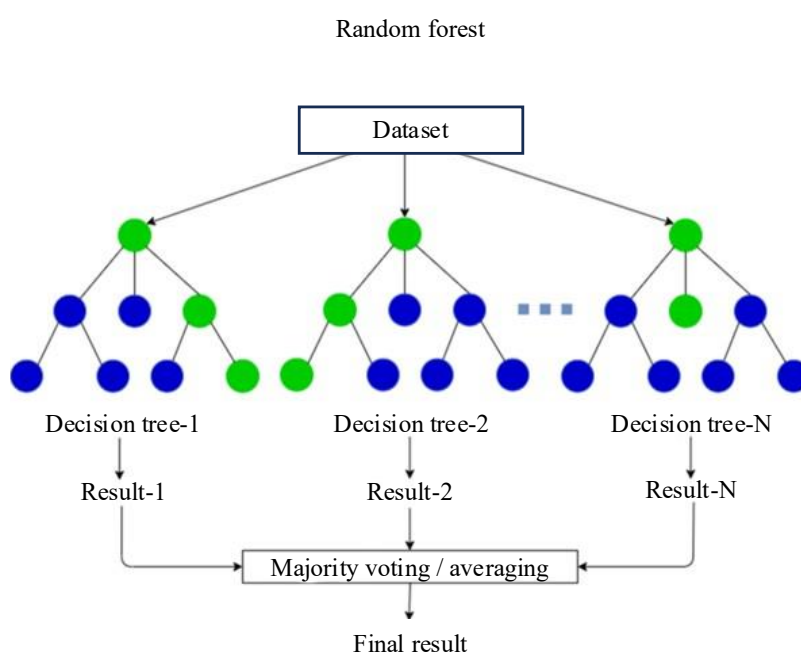


Figure 1. Random forest model.

Equations

For Regression:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x)$$

For Classification:

$$\hat{y} = \text{mode}\{T_1(x), T_2(x), \dots, T_N(x)\}$$

Audience Segmentation

The practice of breaking down a big and varied target market into smaller, easier-to-manage groups according to shared traits like demographics, interests, behaviors, or geographic location is known as audience segmentation. Businesses may use this segmentation to craft more individualized and customized marketing messages that are more likely to connect with each group, enhancing consumer engagement and raising conversion rates. By using audience segmentation, marketers may maximize the efficacy of campaigns by delivering highly relevant information to particular segments of their audience rather than using a one-size-fits-all strategy. By analyzing past data to find organic clusters or groups within the audience based on their reactions to prior ads, machine learning techniques such as Random Forest can improve segmentation. Businesses may use this information to determine which demographics are most likely to interact with particular kinds of content, resulting in more focused and effective marketing campaigns. Moreover, campaigns can use dynamic segmentation, which adjusts in response to real-time behavior data, to improve campaign results and constantly improve targeting.

RoI Prediction

Predicting return on investment (ROI) is a crucial component of campaign strategy and evaluation. It is the process of calculating how much money a company can make from a marketing campaign in relation to how much it will cost. Businesses can prioritize high-performing campaigns, distribute resources efficiently, and pinpoint areas for development when they can accurately forecast return on investment. In marketing, return on investment (ROI) is usually computed by dividing campaign income by campaign expenses, including advertising, production, and staff. Making data-driven judgments regarding future tactics and assessing the effectiveness of marketing initiatives depend heavily on the capacity to anticipate ROI. By examining past data, finding trends, and taking into account several campaign elements (such as the channel utilized, target audience, and ad kind), Random Forest algorithms can help estimate return on investment. Businesses may make better judgments about where to spend their marketing budget by using ROI prediction, which guarantees that money is going to the most profitable and successful tactics. Additionally, companies may use ROI prediction to make real-time campaign tweaks to boost performance and increase profits.

A/B Testing

A/B testing, also known as split testing, is a statistical method used to compare two or more variations of a campaign, product, or strategy to determine which performs better based on a defined metric. It is widely used in marketing, web optimization, and product development to make data-driven decisions. The process involves dividing users or customers into different groups, where each group is exposed to a different version of the campaign or product. Performance metrics, such as conversion rates, click-through rates, or ROI, are then measured and compared to identify the most effective variation. Statistical tests, such as t-tests or ANOVA, are applied to determine whether the observed differences are significant or due to random chance.

In this project, A/B testing is used to evaluate the effectiveness of different marketing campaigns. We analyze key metrics like conversion rates and ROI to determine which campaign performs best. By applying statistical tests, we ensure that any observed differences in performance are meaningful. To enhance the accuracy of our analysis, we also integrate machine learning models, such as Random Forest and Gradient Boosting, which predict campaign success based on historical data. By combining A/B testing with predictive analytics, we can make more informed marketing decisions. If both approaches agree on the best campaign, it reinforces our confidence in the results. However, if they

provide conflicting outcomes, further investigation is required to refine our conclusions and improve future campaigns.

Data-Driven Insights

Actionable results from the statistical analysis and machine learning models of massive datasets are known as data-driven insights. Businesses may depend on objective, evidence-based conclusions in marketing instead of intuition and subjective judgment thanks to data-driven insights. Businesses can find important patterns and trends that guide strategy creation by examining past campaign data, consumer behavior, and other success measures. For instance, data-driven insights may show the best ways to reach target audiences, the ideal times to interact, or the kinds of campaigns that work well for particular demographics. Because machine learning algorithms like Random Forest can evaluate large volumes of data and reveal intricate correlations between variables that might not be immediately obvious, they are especially good at producing these discoveries. Marketers may use these facts to inform choices like redistributing resources, modifying targeting tactics, or modifying message. Additionally, data-driven insights may offer an ongoing feedback loop that enables companies to track campaign effectiveness and make real-time strategy adjustments, guaranteeing that marketing initiatives stay effective and relevant. In the end, data-driven insights enable companies to improve consumer experiences, strengthen marketing tactics, and increase campaign performance.

LITERATURE REVIEW

According to Shihab *et al.*, this system's customer segmentation based on RFM (Recency, Frequency & Monetary) factors is quite popular these days for dividing up the clientele depending on their traits and actions. To improve clustering outcomes, a number of clustering techniques have been used to divide clients into groups [1]. This article compares the agglomerative, k-means, and improved version of k-means approaches for market segmentation based on RFM. In contrast to k-means and improved versions of k-means clustering, the experimental results demonstrate that agglomerative clustering requires a significant amount of processing time for big datasets. Agglomerative clustering does, however, have the important benefit of not requiring the initial designation of the entire number of clusters. The segmentation results also show that the advanced version of k-means can successfully boost the speed of clustering by reducing the overall running time by 27.8 and 97.8%, respectively, when compared to regular k-means and agglomerative clustering. In terms of intra- and inter-cluster distance, it can also yield superior clustering outcomes compared to conventional k-means. Our suggested clustering-based customer segmentation approach makes use of an enhanced version of k-means, which lowers the total time complexity and works rather well with big datasets. Currently, the RFM model is widely used for customer segmentation. Large datasets are typically used as input for this strategy. Nevertheless, not all clustering algorithms can effectively divide clients into several categories. Three clustering techniques were used in this work: agglomerative clustering, k-means, and advanced k-means. The findings of the experiment demonstrate that agglomerative clustering is completely impractical due to its lengthy execution time. However, as compared to regular k-means and agglomerative clustering, the advanced k-means reduces the overall running time by 27.8 and 97.8%, respectively, and outperforms standard k-means in terms of clustering speed and quality. Advanced k-means is more practical than the other two clustering techniques, according on the analysis above. Therefore, our suggested model for RFM-based consumer segmentation is more practical and efficient in terms of speed and clustering quality than previous baseline techniques since it uses advanced k-means clustering. Therefore, our suggested model employing the sophisticated k means method is more practical and efficient for the RFM-based consumer segmentation technique. Using RFE and RFD variables where E and D stand for engagement and duration, respectively we will attempt to apply consumer segmentation in the future. Online businesses are most affected by these keywords [1].

According to Smaili and Hachimi, this method suggests that one of the cornerstones of an effective advertising campaign is customer segmentation [2]. In the process of promoting new products, marketers place a lot of emphasis on this flag ship period. A productive customer marketing strategy will result from effective "customer targeting", which is a prerequisite for successful segmentation.

Customer segmentation has been the subject of several studies that have used the well-known Recency, Frequency, and Monetary model to unsupervised machine learning methods like K-Means. Because it ignores other crucial criteria based on the application field, the model is inadequate. Diversity "D", which refers to the variety of goods a particular client purchases, is the fourth parameter that we included to the model in this study. In the retail industry, RFM-D-based segmentation is used to identify customer behavior patterns. The suggested methodology improves the accuracy of consumer behavior predictions; businesses can forecast which customers would react favorably. In a retail marketing environment, the primary goal shared by all businesses is to maximize return on investment. This can be achieved through acquisition, influencing and attracting new customers, or retention through prospecting for new revenue or making new offers to current customers. Of all customers, only one fifth contributes more to the company's revenue than the other customers. Therefore, we will concentrate on client retention through effective segmentation in this post. To maximize return on investment, that segmentation will group customers and provide them with new items. The objective is to provide a prediction that pinpoints the clients who are most likely to embrace the company's new offerings. It has been established in earlier studies that segmentation using the RFM model is less effective than other approaches such CHAID segmentation [8], segmentation based on CLV, or machine learning-based segmentation [8]. By examining the high association between diversity and the previous model parameters, this research first demonstrates how adding the diversity component greatly enhances the RFM model. Subsequently, the RFM-D model segmentation produced distinct clusters that were of higher quality than the RFM, particularly in terms of the total amount spent by clusters (Figure 1) and the forecast of clients who would accept new offers. Marketers will target customers more successfully and profitably with this segmentation, which enhances the ability to forecast consumer behavior [2].

In this method, Alsayat suggested that user-generated content and big social data have become crucial sources of rich and fast information for identifying trends in consumer behavior [3]. Using user-generated material to demonstrate client happiness has been a major problem in business, particularly in the travel and hotel industry. Numerous research on customer satisfaction have used quantitative survey methods. However, employing massive social data in the form of electronic word-of-mouth (eWOM) to disclose consumer happiness might be a useful method to better understand what customers want. In order to evaluate massive social data on tourists' lodging decisions in Mecca, Saudi Arabia, want to create a hybrid methodology based on supervised learning, text mining, and segmentation machine learning techniques. Developing the hybrid technique using k-means, latent Dirichlet allocation (LDA), and support vector regression with sequential minimum optimization (SMO) [4]. Information from TripAdvisor reviews of hotels in Mecca. Travelers' contentment is shown for each portion of the data based on their online hotel reviews. The findings demonstrate the method's efficacy for traveler segmentation in Mecca hotels and huge social data analysis. The findings are examined, and a number of suggestions and tactics are offered to hotel managers in order to raise the caliber of their services and raise client happiness. This study used k-means, latent Dirichlet allocation (LDA), and support vector regression with sequential minimum optimization to provide a novel technique to consumer segmentation [5]. TripAdvisor provided the information on internet reviews of hotels in Mecca. Traveler satisfaction was shown for each part of the data after it was segregated using k-means. To analyze textual data and identify the characteristics of pleasure for each section, the LDA approach was employed. In order to anticipate customer satisfaction (overall ratings) based on a number of performance quality variables, SVR-SMO was finally implemented in each sector. The RMSE, MAE, and R^2 were used to gauge how effective the suggested approach was. The assessments' outcomes were then contrasted with those derived from the decision tree, MLR, and ANN techniques. The findings showed that the k-means and SVR-SMO approach had a greater prediction accuracy than the other approaches. There are several limitations to this study that need be addressed in further research. TripAdvisor was utilized to gather data for this study [3].

RELATED WORK

Using a pre-processed dataset of 495,478 customers, study examines consumer behavior in the UK retail sector. It uses a variety of techniques, such as Box-Cox transformation, RFM (Recency,

Frequency, Monetary) segmentation, and temporal analysis, to forecast customer lifetime value (CLV). Several supervised machine learning models were used, including Random Forest, AdaBoost, Extra Trees, LightGBM (LGBM), and XGBoost. The greatest performance in predicting CLV was demonstrated by XGB Classifier and Extra Trees. The study sets a new benchmark in retail analytics by providing a scalable, effective technique for forecasting consumer behavior and highlighting the importance of RFM segmentation in understanding customer behavior.

METHODOLOGY

The suggested system analyzes and improves marketing campaign effectiveness using machine learning techniques, particularly the Random Forest algorithm. A large dataset with attributes like Campaign ID, Company, Campaign Type, Target Audience, Duration, Channel Used, Conversion Rate, Acquisition Cost, ROI, and Location is processed by the system. To ensure that the model receives high-quality input, it starts by preprocessing the data to address missing values and inconsistencies. In order to identify the key elements that lead to campaign success, the Random Forest method is used to model connections within the data. The algorithm forecasts conversion rates and return on investment by examining these variables, and it also identifies the most cost-effective channels and ideal campaign kinds. Furthermore, it divides consumers into groups according to their reaction and engagement habits, enabling more specialized and focused marketing tactics [6]. In addition to offering useful information for decision-making, this clever method enables companies to maximize marketing expenditures, improve resource allocation, and execute campaigns more effectively.

Load Data

The marketing campaign data must be imported into the system via the Load Dataset Module. It guarantees that the data is correctly put into memory for additional processing and supports a number of data types, including CSV, Excel, and database connections. By extracting pertinent data elements that will be utilized for analysis, including Campaign ID, Company name, Campaign Type, Target Audience, Duration, Channel utilized, Conversion Rate, Acquisition Cost, ROI, and Location, this module is crucial for system setup [7]. It guarantees that the dataset is available for use in the pipeline's later phases of data processing.

Data Preprocessing

The dataset must be cleaned and prepared for analysis via the Data Preprocessing Module. It manages typical data issues such as irrelevant characteristics, missing values, and inconsistencies [8]. This module employs methods such as encoding to transform categorical variables into numerical representations, normalization to scale numerical features, and imputation to fill in missing data. This guarantees that the data is formatted appropriately for machine learning models, enhancing the precision and dependability of further research.

Feature Analysis

This module finds and chooses the most pertinent characteristics that have a big impact on marketing campaign results. It highlights important characteristics like Campaign Type, Target Audience, Channel Used, and ROI using statistical methodologies and feature selection procedures [9]. By concentrating on these crucial elements, the method guarantees that it catches the most significant patterns in the data while also simplifying the model. Enhancing the model's effectiveness and interpretability requires this step.

Modeling and Prediction Module

The Random Forest approach is used by the Modeling and Prediction Module to examine the connections among the dataset's different attributes. With this ensemble approach, many decision trees are built, each of which bases its predictions on a distinct feature of the data. The Random Forest algorithm compiles the data to produce precise forecasts of important outcomes, such as ROI and conversion rates [10]. This module is essential to the system's capacity to estimate campaign performance since it captures intricate feature interactions and guarantees accurate forecasts.

Audience Segmentation Module

The target audience is divided into categories by the Audience Segmentation Module according to characteristics including response patterns, behavior, engagement levels, and demographics. This module makes it possible to create customized marketing plans that are more likely to connect with each category by putting like consumers together. This module assists companies in concentrating their efforts on the most promising target groups in order to optimize their marketing effect. Customized campaigns have a greater conversion rate, leading to increased customer satisfaction, stronger brand loyalty, and higher overall revenue generation.

ALGORITHM DETAILS

Marketing initiatives, which engage targeted audiences through customized methods across several media, are crucial for propelling corporate success. Campaign type, duration, target audience, and channel performance are some of the variables that affect these campaigns' efficacy. These variables all work together to affect metrics like ROI, acquisition costs, and conversion rates.

The Random Forest algorithm, an ensemble learning method that creates many decision trees during the training stage, is used in this work. To provide variation across the trees, each decision tree is built using a random subset of the characteristics and data. By averaging or voting on the predictions from separate trees, the algorithm uses strategies like bagging (bootstrap aggregating) to limit overfitting and reduce variation. The Random Forest algorithm can describe intricate patterns and relationships in the data thanks to this organized approach, producing predictions that are precise and trustworthy. The insights gained from this method help to optimize decision-making processes and comprehend the links between qualities.

RESULT ANALYSIS

The results study shows that using machine learning techniques, specifically, the Random Forest algorithm, to assess and improve marketing campaigns is effective. The model finds intricate patterns and relationships between these variables by analyzing a dataset that includes characteristics like campaign kind, target audience, length, channels used, conversion rates, acquisition costs, ROI, and geography.

The method efficiently establishes each attribute's relative relevance, emphasizing the components that have the biggest effects on campaign results. The model's predictive powers enable the assessment of anticipated conversion rates and return on investment across different campaign setups. Important information on ideal campaign lengths, appropriate target segments, and the efficacy of various marketing platforms is revealed by this study. Furthermore, the study makes it easier to divide audiences into discrete groups according to similar traits, which makes it possible to spot distinctive trends in how they react to different advertising tactics. By identifying cost-effective routes that maximize returns while lowering acquisition costs, the model's output also offers a deeper knowledge of cost dynamics. Additionally, it reveals patterns that guide strategic planning, such determining the kinds of campaigns that appeal best to particular target segments. Utilizing these results, the research provides practical suggestions for creating campaigns that are more effective and memorable, with a focus on resource optimization and strategy improvement. These findings highlight how data-driven strategies may improve the analytical skills necessary for intricate campaign management decision-making processes (Figure 2).

Table 1 represents the accuracy level between existing and proposed work.

Table 1. Comparison table.

Algorithm	Accuracy
XGBOOST	70
RF	85

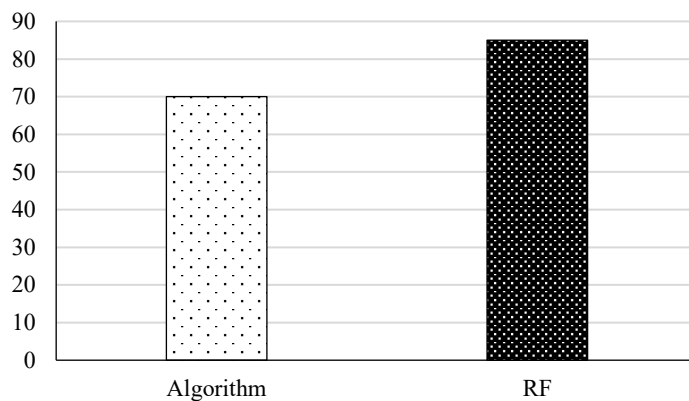


Figure 2. Comparison graph.

Figure 2 shows the graphical representation of the algorithms of XGBOOST and RF.

CONCLUSION

In conclusion, by offering exact forecasts and actionable insights, the suggested method makes use of machine learning, more especially the Random Forest algorithm, to improve the effectiveness of marketing efforts. The system efficiently analyzes complicated campaign data to maximize decision-making using a systematic methodology that includes data pre-treatment, feature analysis, modeling, audience segmentation, and performance evaluation. Businesses may enhance overall campaign effectiveness and allocate resources more effectively by identifying the important aspects that influence campaign success, such as the most cost-effective channels and the most receptive target groups. Proactive strategy modifications are made possible by the system's capacity to forecast outcomes like conversion rates and ROI, which eventually maximizes return on investment. Businesses in a variety of sectors may benefit greatly from the system's scalability and adaptability, which promotes more focused and successful marketing campaigns. Future marketing efforts will be more successful and economical as the system's accuracy and impact continue to improve with ongoing feedback and real-world data.

Future Work

Future development of the suggested system may concentrate on enhancing its functionality by adding more machine learning models and cutting-edge methods to improve prediction accuracy even more. Investigating deep learning techniques like gradient boosting or neural networks may yield more reliable models for managing intricate and big datasets. Furthermore, using real-time data streams may allow for dynamic modifications to marketing plans in response to real-time campaign results, giving companies faster and more accurate insights.

In order to increase audience segmentation and develop even more individualized marketing strategies, it could also be beneficial to use more detailed consumer data, such as behavioral data from online interactions, social media activity, or customer feedback. Additionally, the system may be expanded to accommodate multi-channel campaigns, which would enable companies to assess and improve integrated marketing initiatives across many platforms at the same time. Last but not least, creating an intuitive dashboard and user interface would increase the system's usability for marketers, allowing them to quickly analyze findings and take action without requiring extensive technical knowledge. These developments would enhance the system's scalability and efficacy, making it a useful instrument for companies across a range of sectors.

REFERENCES

1. Shihab SH, Afroge S, Mishu SZ. RFM based market segmentation approach using advanced k-means and agglomerative clustering: a comparative study. In 2019 IEEE International conference on electrical, computer and communication engineering (ECCE). 2019 Feb 7; 1–4.

2. Smaili MY, Hachimi H. New RFM-D classification model for improving customer analysis and response prediction. *Ain Shams Eng J.* 2023 Dec 1; 14(12): 102254.
3. Alsayat A. Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. *Neural Comput Appl.* 2023 Feb; 35(6): 4701–22.
4. Bratina D, Faganel A. Using supervised machine learning methods for RFM segmentation: a casino direct marketing communication case. *Market-Tržište.* 2023 Jun 9; 35(1): 7–22.
5. Bahari TF, Elayidom MS. An efficient CRM-data mining framework for the prediction of customer behaviour. *Procedia Comput Sci.* 2015 Jan 1; 46: 725–31.
6. Lewaa I, Hafez MS, Ismail MA. Data integration using statistical matching techniques: A review. *Stat J IAOS.* 2021 Nov; 37(4): 1391–410.
7. Lewaa I, Hafez MS, Ismail MA. Mixed Statistical Matching Approaches Using a Latent Class Model: Simulation Studies. *J Stat Appl Probab.* 2023; 12(1): 247–65.
8. Rahim MA, Mushafiq M, Khan S, Arain ZA. RFM-based repurchase behavior for customer classification and segmentation. *J Retail Consum Serv.* 2021 Jul 1; 61: 102566.
9. Arefin S, Parvez R, Ahmed T, Jahin F, Sumaiya F, Ahsan M. Enhancing bank consumer retention: a comprehensive analysis of churn prediction using machine-learning and deep learning techniques. In *International Conference on Advances in Information Communication Technology & Computing*. Singapore: Springer Nature Singapore; 2024 Apr 29; 265–281.
10. He X, Li C. The research and application of customer segmentation on e-commerce websites. In *2016 IEEE 6th International Conference on Digital Home (ICDH)*. 2016 Dec 2; 203–208.