Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD30299

Application of Mathematical Statistical Models in Big Data User Behavior Analysis

Yuxiang He

Dana and David Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, USA
heyuxian@usc.edu

Abstract. In today's world of digital business and interaction, more and more companies rely on data-driven strategies to know, predict, and guide user behavior. In this study, I do not only aim to operationalize these models using a large dataset but also to assess the relative strengths, limitations, and complementary nature of these models. In addition, referencing recent studies in academic and industry on the impact of COVID on behaviors, how to create fintech-specific personalization, and sequential behavioral modeling, the paper ties itself to the large body of related works. This paper uses RFM and Logistic Regression statistical models to analyze user behavior in big data scenarios, identifying patterns in a real-world dataset from Kaggle's "Customer Personality Analysis," integrating marketing data like demographics and campaign responsiveness.

Keywords: RFM Model, Customer Segmentation, Logistic Regression, E-Commerce, Consumer Behavior

1. Introduction

In today's world of digital business and interaction, more and more companies rely on data-driven strategies to know, predict, and guide user behavior. With behavioral traces escalating exponentially, the ability to transform these raw behavioral traces into actionable insights has risen as a competitive benefit and a scientific problem. In this context, mathematical statistical models provide an important role in providing systematic frameworks to extract patterns from the data and to understand the data patterns, to be able to predict and prescribe. This paper employs two general-purpose statistical models, namely the RFM (Recency, Frequency, Monetary) model and Logistic Regression to analyze user behavior in big data scenarios. The study uses these models to discover a pattern in a real-world dataset from the open-source Kaggle repository named "Customer Personality Analysis". This dataset integrates comprehensive marketing-related information such as customer demographics, purchasing patterns, and campaign responsiveness in this dataset. So, such a comprehensive dataset is well-suited to evaluating the efficacy of segmentation and classification models in marketing analytics. For quite a while now, the RFM model has been one of the most popular methods for customer segmentation in the direct marketing context. This model categorizes user purchasing behavior into three measurable dimensions: how recently a customer has made a purchase (Recency), how frequent their purchasing is (Frequency), and how much they spend

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD30299

(Monetary). Scoring each user along these axes allows for easy customer grouping and the efficient allocation of marketing resources in a business. The study then analyze the data based on the evolution of the distributions of RFM score and analyze the aggregate of the RFM score distributions with a focus on the groupings of its behavioral clusters, following methodological frameworks from the previous few analyses [1,2].

However, Logistic Regression presents a solid probabilistic model for performing binary classification tasks. “Its key strength lies in estimating the likelihood of user actions such as conversion or churn – conditional on such features as past interaction history, interaction with past campaigns, and demographics. So, in line with what is done in e-retail studies whose focus is predicting customer response to promotional efforts, this study utilizes Logistic Regression for this predictive task [3]. The study aims not only to operationalize these models using a large dataset but also to assess the relative strengths, limitations, and complementary nature of these models. In addition, referencing recent studies in academic and industry on the impact of COVID on behaviors [4,5], how to create fintech-specific personalization (Attota, 2024), and sequential behavioral modeling, the paper positions itself within the broader body of related literature [6].

2. Methodology

In this section, the RFM model as well as the Logistic Regression model were implemented using Python and with the marketing_campaign.csv dataset. The pandas, NumPy, scikit-learn, matplotlib, and seaborn libraries were used in the analysis.

2.1. RFM model implementation

RFM model takes multiple product category columns to calculate recency (days since last purchase), frequency (total purchases), and monetary (total spending). To enable cross-metric comparison, MinMaxScaler was applied for normalization, followed by quantile-based binning was performed to assign 1 to 5 scores to each metric. The resulting total RFM score is the sum of the three RFM scores (3 to 15).



Figure 1. Customer distribution by RFM score

DOI: 10.54254/2755-2721/2025.LD30299

Table 1. RFM summary table

RFM_Score	Count	Avg_Recency	Avg_Frequency	Avg_Monetary
3	77	0.892824	0.115998	0.011456
4	118	0.785739	0.138867	0.016714
5	174	0.662777	0.154258	0.024834
6	199	0.504644	0.167314	0.030268
7	218	0.429756	0.201939	0.052022
8	195	0.450557	0.264802	0.104025
9	224	0.498151	0.330763	0.200105
10	239	0.614936	0.435527	0.344888
11	265	0.526587	0.476244	0.410767
12	188	0.375833	0.484043	0.439864
13	167	0.286760	0.523272	0.514877
14	130	0.150738	0.555245	0.510650
15	22	0.095041	0.608471	0.611418

Figure 1 (Customer Distribution by RFM Score) shows the customer segmentation curve, peaking around scores 10–11. Table 1 summarizes average recency, frequency, and monetary value across score levels. These outputs enable direct targeting of high-score segments with retention campaigns [7].

2.2. Logistic regression implementation

The campaign response likelihood is predicted using a logistic regression model from demographic and behavioral features. On the training set, the binary target variable (Response) was encoded and features were normalized. Since the columns 'ID', 'Dt_Customer', and 'Z_CostContact' are irrelevant and features required standardization, the study used StandardScaler to standardize the relevant columns and removed all irrelevant variables [7]. The dataset was further partitioned 80/20 into training and test sets.

Table 2. Classification report

	precision	recall	f1-score	support
0	0.90	0.96	0.93	382
1	0.57	0.34	0.42	62
accuracy			0.87	444
macro avg	0.73	0.65	0.68	444
weighted avg	0.85	0.87	0.86	444

The summary of results included an accuracy of 87.2% and an ROC AUC score of 0.893, confirming its strong generalization capability. Table 2 shows the precision, recall, and F1 of each class of the model. In particular, the model is good at finding non responders which is very useful for spending optimization. Together, RFM constitutes a two-stage pipeline: customer segmentation

DOI: 10.54254/2755-2721/2025.LD30299

using the RFM model, followed by Logistic Regression for personalized targeting within each segment..

3. Overview

This study analyzes the segmentation and prediction capabilities of the RFM and Logistic Regression models in user behavior analysis, and demonstrates their respective advantages and complementary relationships through an analytical framework. While Logistic Regression can be used to identify the most probable responders to targeted campaigns, the RFM model is great for classifying customer value. RFM scores distribution shows central tendency in the middle of the scoring range and the most important for users scores are: 10 and 11. Usually, these levels of score refer to the users who have made recent purchases, higher frequency, and large monetary value, which are similar to marketing high priority targets. In contrast, lower RFM scores indicate customers who require re-engagement or should be excluded from premium campaign resources. At the same time, logistic regression provided insight into how various user attributes are related to responsiveness to a campaign. More positive thinking fell along those same lines, as younger customers, those with prior engagement in campaigns, and those with lower spend segments tended to respond more favorably—further supported by pandemic-era behavior change under stress and isolation. The reliability of the model in binary classification scenarios, even with class imbalance, is further proven by the model's ROC AUC of 0.893.

The combination of these models gives a two-step strategic framework that first segments the users of interest using RFM, and then identifies within each segment those to prioritize using logistic regression. The combination of these two objectives enables the formation of macro-level strategic groupings and micro-level tactical activation. These insights resonate with SeqRFM, as segmentation combined with sequential modeling yields optimal marketing outcomes. Finally, the integrated dual-model structure covers the customer journey from the static value prediction to the dynamic intent prediction, further providing great potential to support the automated personalization of an array of products under a large-scale customer database.

4. Discussion

In the era of big data, the analysis of user behavior has become increasingly refined, and this study demonstrates that classical statistical tools can be adapted into solutions tailored to address contemporary user behavior challenges. The strengths and limitations of all models will be specific to each individual model. It has transparency and operational efficiency offered by the RFM model. It empowers marketers to separate customers with a view to their buying conduct using a lightweight scoring framework. RFM scores are not only very simple to compute but also to interpret; thereby RFM becomes an attractive proposition for companies that do not have the data science infrastructure. Nevertheless, being static and with aggregate data, it hid the nuances of behavior. E-commerce Marketing claims that it also assumes recency, frequency, and monetary value equally contribute to customer value, an assumption that may not hold in domains.

However, logistic regression takes into account several predictive variables and can model interaction effects. As a customer-centric model, it estimates the probability of a specific event, in this case of campaign acceptance, and based on customer-specific inputs. Because of this flexibility, it can adjust to a user's behavior changes in the presence of external factors like seasonality, policy change, or economic shocks. Despite this, logistic regression is limited as well. However, it is restrictive, since it assumes linearity in log-odds. Moreover, and its effectiveness diminishes when

DOI: 10.54254/2755-2721/2025.LD30299

input features exhibit collinearity or the dataset suffers from severe class imbalance. Model interpretability is very important. Business Stakeholders may communicate RFM scores to the system and integrate them directly into CRM systems. Although logistic regression's results are explainable via coefficients and odds ratios, the practitioner needs to know more statistically to interpret them meaningfully. In environments like fintech or healthcare, where transparency is prioritized over statistical rigor, RFM may be considered more attractive than logistic regression [3]. The first and quite innovative approach is a hybrid one, where RFM is first used to segment customers, then logistic regression is applied to each segment. This layered modeling methodology aligns well with both dynamic campaign optimization frameworks and audience prioritization in real-time marketing environments [5]. It also facilitates differential promotion of incentives, allowing premium promotions to consumers with high probability and high RFM scores, while deprioritizing consumers with low right scores. This is, in fact, at least one dual-use case consistent with best-of-breed customer lifecycle management. Early funnel intervention support like engagement or loyalty initiatives are part of RFM segmentation. End-of-funnel strategies such as cross-selling and personalized discounting are supported by logistic regression. Within a real-time marketing stack, this integration can improve ROI deliverables by aligning spending of budget with the probability of realizing expectations of returns.

5. Conclusion and future directions

This study shows how the classical statistical models, RFM and Logistic Regression, can be utilized to learn and optimize customer behavior in the big data realm. Behavioral segmentation was conducted using the RFM modeling based on purchase history and logistic regression helped in making probabilistic predictions of campaign response. Even though they are old paradigms, these models remain relevant in modern, data-rich environments when applied appropriately. In line with the research conducted, the project concludes that RFM is of worth in segmenting users into clear clusters and identifying top-tier customers that contribute incompletely to revenue. At the same time, logistic regression provided reliable approximations to the probability estimates which marketers can use to hone in on campaign targeting and lower acquisition costs. The results of these findings have shown to be in sync with practical as well as academic evaluations in e-commerce as well as fintech settings.

The future work would be the fusion of RFM with other effective classification methods like decision trees, gradient boosting machines, or neural networks. However, these models can give better accuracy and can accommodate nonlinear interactions better but have to pay for the lack of interpretability. Another way is to go beyond static to sequential RFM, using SeqRFM for instance, to update customer value scores in real time based on each new behavioral signal. Similarly, fairness and bias mitigation also have to form part of model governance. Historical biased data could be used to train logistic regression models which might enhance systemic inequities. Feature influence can be evaluated using transparency tools such as SHAP or LIME to avoid nonequitable decisions. Finally, such systems can automate user experiences at scale, they are personalized, based on these models. Future implementations should work towards building full-stack pipelines that ingest behavioral signals, perform RFM scoring, update response probabilities, as well as seamlessly deploy the interventions. In essence, such systems will define the standards for intelligent, ethical, and profitable customer engagement.

DOI: 10.54254/2755-2721/2025.LD30299

References

[1] Alparslan, H. A. K. A. N., Turgay, S., & Yilmaz, R. (2024). Utilizing logistic regression for analyzing customer behavior in an e-retail company. Financial Engineering, 2, 116-125.
[2] Attota, S. Y., & Chowdary, D. (2024). Optimizing fintech marketing: A comparative study of logistic regression and XGBoost. arXiv Preprint arXiv: 2412.16333.
[3] S (2024) "SeqRFM: Fast RFM Analysis in Sequence Data." arXiv preprint arXiv: 2411.05317 .
[4] Anitha, Palaksha, and Malini M. Patil.(2022) "RFM model for customer purchase behavior using K-Means algorithm." Journal of King Saud University-Computer and Information Sciences 34.5: 1785-1792.
[5] Chitanand, P. (2025). RFM analysis: Understand your customers and segmentation. Express Analytics.https: //www.expressanalytics.com/blog/rfm-analysis-customer-segmentation/
[6] Truong, D., & Truong, M. D. (2022, July). How do customers change their purchasing behaviors during the COVID-19 pandemic? Journal of Retailing and Consumer Services, 67, 102963. https: //doi.org/10.1016/j.jretconser.2022.102963
[7] OmniSegment. (n.d.). E-commerce marketing: Identify high-value customers with the RFM model. https: //omnisegment.com/ecommerce-marketing-identify-high-value-customers-with-the-rfm-model/