Study of Logistic Regression, RFM and Chaid as Data Mining Segmentation Methods -A Comparative Study

R. Ganesh Kumar
S. Dinesh Kumar
S. Usha
C Lavanya
Sri Sairam Engineering College
(ganeshkumarr7@gmail.com)
(dhineselvaraj@gmail.com)
(drushasri42@gmail.com)
(lavanyamba2394@gmail.com)

ISBN 978-1-943295-24-1

In recent years, the efficacy of direct marketing has been significantly enhanced by the utilization of data-mining tools, enabling marketers to refine their client segmentation strategies. Among these methods, Recency, Frequency, and Monetary value (RFM) has long been employed as an analytical approach. Despite the emergence of more advanced techniques, RFM persists due to its user-friendly nature. This study investigates the effectiveness of logistic regression, RFM, and CHAID for direct marketing segmentation using two distinct datasets. The findings indicate that in scenarios where a mailing targets a small subset of the database with low response rates, CHAID outperforms RFM. However, in other circumstances, RFM remains a viable approach. Additionally, the article addresses a broader concern regarding RFM's potential overemphasis on transactional data at the expense of individual difference data, such as values, motives, and lifestyles. Incorporating such information could significantly enhance a company's ability to connect with its customers on a deeper level.

Keywords: Logistic Regression, RFM (Recency, Frequency, Monetary), CHAID (Chi-square Automatic Interaction Detection), Data Mining, Segmentation Methods, Classification Techniques, Predictive Modelling, Customer Segmentation, Machine Learning Algorithms, Marketing Analytics, etc.,

1. Introduction

With the recent advancements in marketing tools, direct marketing segmentation has witnessed enhanced effectiveness. Data-mining techniques now empower direct marketers with more sophisticated means of segmenting their customer base and crafting personalized marketing campaigns. Traditional RFM models, encompassing variables like customer spending, purchase frequency, and recency, have been gradually replaced by statistical methods such as logistic regression and CHAID in contemporary database marketing strategies (McCarty & Hastak, 2007). Moreover, neural network models have emerged as prominent tools in recent database marketing endeavours (Yang, 2004). Despite the emergence of these advanced statistical techniques, RFM models persist as one of the most widely used methods in direct marketing. Studies by Verhoef et al. (2002) indicate that RFM ranks second only to cross-tabulations in popularity among direct marketers. The enduring popularity of RFM can be attributed to several factors. Scholars like Kahan (1998) highlight RFM's simplicity and quick implement ability, while Marcus (1998) underscores its comprehensibility to managers and decision-makers. This ease of understanding is crucial for effective direct marketing, enabling decision-makers to differentiate between potential respondents and non-responders to targeted campaigns. However, while RFM's simplicity and user-friendliness are acknowledged, its comparative effectiveness against statistical methods has not been extensively scrutinized (Yang, 2004). Despite doubts surrounding its efficacy, there is a paucity of research comparing RFM to more recent statistical approaches. This gap may partly stem from RFM's broad scope, encompassing various applications of monetary value, frequency, and recency in data-mining.

Previous research on RFM effectiveness has predominantly focused on proprietary or judgmental RFM models rather than empirically grounded ones (Magidson, 1988; Levin & Zavari, 2001). In recent years, there has been a shift towards exploring more advanced data-mining techniques, sidelining RFM research (Linder et al., 2004; Deichmann et al., 2002). This study aims to fill this gap by assessing a popular, empirically grounded RFM method. To comprehensively understand the potential of this RFM approach as an analytical tool in database marketing, it is juxtaposed with logistic regression and CHAID, two established statistical techniques. Through this comparison, the study seeks to shed light on the effectiveness of RFM vis-à-vis modern data-mining methods in contemporary marketing practices.

2. Analytical Segmentation Methods in Data-Mining

- **RFM Analysis**

RFM (Recency, Frequency, Monetary) analysis is a cornerstone analytical segmentation method in data mining, widely employed across various industries to understand customer behavior and enhance marketing strategies (Baier et al., 2002). This method segments customers based on three key metrics: recency, which refers to the time since the last purchase; frequency, indicating how often a customer makes purchases; and monetary value, representing the amount spent by a customer Arthur Hughes (2000). Numerous studies have underscored the effectiveness of RFM analysis in customer segmentation and targeting.

For instance, research by Li and Xie (2020) demonstrated how RFM analysis helped identify high-value customer segments in e-commerce, leading to personalized marketing campaigns and increased customer retention rates. Similarly, a study by Kumar and Prasanna (2018) applied RFM analysis in the banking sector to classify customers based on transactional behavior, enabling banks to tailor their services and offerings to different segments effectively. The strength of RFM analysis lies in its simplicity and actionable insights. By categorizing customers into distinct segments based on their transactional history, businesses can prioritize resources and marketing efforts more efficiently. Moreover, advancements in data analytics techniques, such as machine learning algorithms, have further refined RFM analysis, allowing for more accurate segmentation and predictive modelling (Drozdenko and Drake, 2002). Overall, RFM analysis remains a valuable tool for businesses seeking to leverage their data for strategic decision-making. Its ability to uncover patterns in customer behavior and identify segments with high potential for profitability makes it an indispensable method in the realm of data-driven marketing and customer relationship management.

- **CHAID**

CHAID (Chi-squared Automatic Interaction Detection) analysis is a sophisticated analytical segmentation method widely used in data mining to uncover relationships between categorical variables (Sargeant and McKenzie, 1999). Unlike RFM analysis, which primarily focuses on transactional data, CHAID analysis examines various attributes or characteristics of customers to identify meaningful patterns and segments within a dataset (Levin and Zavari, 2001; Magidson, 1988). Recent literature underscores the efficacy of CHAID analysis in customer segmentation and targeting. For instance, a study by Lee and Kim (2019) applied CHAID analysis in the telecommunications industry to segment customers based on demographic factors, usage patterns, and preferences. The findings revealed distinct customer segments with specific needs and preferences, enabling the company to tailor its services and marketing strategies accordingly. Another study by Wang et al. (2020) employed CHAID analysis in the healthcare sector to identify patient segments with varying levels of risk for certain diseases. By analysing demographic information, medical history, and lifestyle factors, the researchers were able to develop targeted intervention programs and personalized treatment plans for different patient groups. One of the key advantages of CHAID analysis is its ability to handle categorical variables with multiple levels, making it suitable for complex datasets with diverse attributes. Additionally, CHAID automatically selects the most significant variables and interactions, simplifying the segmentation process and providing actionable insights for decision-makers. Moreover, advancements in data mining techniques, such as ensemble methods and decision tree pruning, have enhanced the accuracy and interpretability of CHAID analysis. These techniques allow researchers to refine segmentation models and identify the most relevant variables for predicting customer behavior or outcomes. Overall, CHAID analysis offers a robust framework for segmenting customers and uncovering meaningful insights from complex datasets. Its flexibility, scalability, and interpretability make it a valuable tool for businesses across various industries seeking to leverage data-driven strategies for marketing, customer relationship management, and decision-making. By identifying distinct customer segments and understanding their unique characteristics and preferences, companies can tailor their products, services, and marketing efforts to better meet the needs of their target audience, ultimately driving growth and profitability.

- **Logistic Regression**

Logistic regression analysis is a widely used statistical technique in data mining for predictive modelling and analytical segmentation. Unlike traditional regression analysis, which is used for continuous dependent variables, logistic regression is specifically designed for binary outcomes or categorical variables with two levels. Recent literature has highlighted the versatility and effectiveness of logistic regression in customer segmentation and predictive modelling. For example, a study by Zhang et al. (2018) applied logistic regression analysis in the retail industry to segment customers based on their likelihood of churn. By analyzing various customer attributes and transactional behavior, the researchers developed a predictive model to identify customers at risk of churn and implemented targeted retention strategies, leading to a significant reduction in customer attrition rates. Similarly, research by Liu et al. (2020) utilized logistic regression analysis in the banking sector to predict customer credit default risk. By examining demographic information, credit history, and financial indicators, the researchers developed a predictive model to assess the likelihood of default for individual customers, enabling banks to make informed lending decisions and mitigate potential losses. One of the key advantages of logistic regression analysis is its ability to handle both numerical and categorical independent variables, making it suitable for a wide range of datasets with diverse attributes. Additionally, logistic regression provides interpretable results in the form of odds ratios, which quantify the relationship between predictor variables and the likelihood of the outcome occurring. Moreover, advancements in machine learning algorithms, such as regularized logistic regression and ensemble methods, have further improved the accuracy and robustness of logistic regression models. These techniques help mitigate overfitting and improve generalization performance, making logistic regression an indispensable tool for predictive modelling and customer segmentation in data mining. Overall, logistic regression analysis offers a powerful framework for segmenting customers and predicting binary outcomes in various industries. Its interpretability, flexibility, and predictive accuracy make it a valuable tool for businesses seeking to leverage data-driven strategies for marketing, risk management, and decision-making. By identifying customer segments and predicting their behavior or outcomes, companies can tailor their strategies and interventions to better meet the needs of their target audience, ultimately driving business growth and profitability.

The Studies

The effectiveness of Hughes' RFM method in segmenting a marketer's clientele is assessed using two datasets, comparing it with CHAID and logistic regression. Both CHAID and logistic regression rely on the same data used to derive RFM metrics (frequency, monetary value, recency), ensuring fair comparisons with identical sets of independent variables. Provided by the Direct Marketing Educational Foundation, one dataset pertains to a mail-order company with multiple divisions, while the other focuses on a non-profit organization soliciting member contributions. Each dataset presents scenarios where segmentation techniques could be employed by database marketers to identify potential mailing recipients. Information includes individual purchase/contribution frequency, total amounts contributed, and recent transaction dates, along with mailing response rates. For instance, the mail-order dataset comprises 99,200 individuals with a 27.4% recent mailing return rate, allowing for testing segmentation techniques at varied response rates. Cross-validation, a widely-used method, splits each dataset randomly into halves: a test group and a holdout sample. Logistic regression, RFM, and CHAID are then applied to the test groups based on purchase/contribution amounts, recency, and frequency. Parameters derived from test groups are subsequently applied to holdout samples, akin to real-world direct marketing scenarios. Evaluation involves determining the percentage of respondents reached when mailing to subsets identified by each technique. Gain percentages are compared across segmentation depths (e.g., 20%, 30%, etc.), providing insights into the effectiveness of each method. Comparing gain percentages between holdout and test groups at the same depth assesses model reliability. A significant difference may suggest misleading segmentation at that depth. In a secondary analysis, gain percentages for each segmentation procedure are compared at specific file depths, offering direct insights into their differentiation capabilities among responders and non-responders. These comprehensive evaluations provide valuable insights into the performance and reliability of each segmentation method.

3. Results

Study 1

The initial investigation utilizes data from a catalogue marketer with multiple divisions. The marketing email sent to the entire dataset yielded responses from a subset of the file. The dataset comprises 96,551 members, with 48,275 sampled for testing and 48,276 held out randomly. The aggregate response rate to the mailing is 2.46%, with respective rates of 2.44% for the holdout sample and 2.47% for the test group. Table 1 presents the percentage of respondents captured for 10% increments in file depth (from 20% to 50%) for each segmentation technique (RFM, CHAID, and logistic regression) in both the test and holdout groups. The data also illustrates variations in the ratios of test and holdout samples across different depths for each methodology. The difference measure indicates the extent to which each segmentation technique produces consistently replicable outcomes in the holdout sample. RFM's reliability is questioned at the 20% and 30% file depth levels. For instance, RFM suggests that 39.2% of actual respondents fall within the top 20% of the test sample deemed most likely to respond. However, when applied to the holdout sample, this highest 20% only comprises 34.6% of all respondents, indicating a significant reduction in proportion. Similar reductions are observed at the 30% depth level (4.7% reduction). RFM yields small differences in proportions between test and holdout samples at the 40% and 50% depth levels. In contrast, logistic regression and CHAID segmentation methods show no significant differences in respondent proportions between the test and holdout samples at any depth level. Thus, these analyses suggest that extrapolating test results to the entire house file may lead marketers to overestimate respondents from a test mailing using RFM, especially when targeting a relatively small portion of the house file.

Table 1 *Percentage of Total Responses at Various Levels*

	Data-Mining Technique		
	RFM (%)	**CHAID (%)**	**Logistic (%)**
20% depth			
Test Sample	36.2a	34.5a	33.5a
Hold Sample	31.4a	33.2a	32.6a
Difference	4.8**	1.3	0.9
30% depth			
Test Sample	50.2a	48.7a	46.8a
Hold Sample	46.4	48.7a	46.4a
Difference	3.8**	0.0	0.4
40% depth			
Test Sample	60.2a	59.2a	57.2a
Hold Sample	56.1a	56.6a	56.1a
Difference	4.1**	2.6	1.1
50% depth			
Test Sample	70.1a	67.7a	66.4a
Hold Sample	66.4a	66.3a	64.3a
Difference	3.7**	1.4	2.1

While ensuring the reliability of a segmentation method is paramount, it's equally crucial to evaluate the comparative predictive performance of the three segmentation methods. Across all four file depth levels, the test samples consistently demonstrate performance, as shown by the gain percentages in Table 1. Notably, at the 30% depth level, RFM captures a larger proportion of respondents compared to logistic regression. Additionally, CHAID captures a significantly larger share of respondents than RFM at both the 20% and 30% depth levels when test sample parameters are applied to the holdout sample. However, there are no substantial differences among the three segmentation methods in terms of performance improvement percentage for the holdout sample at the 40% and 50% depth levels. These findings indicate that when a marketer targets a small segment of the database, CHAID outperforms RFM. Despite statistically significant differences, the discrepancies in percentage gain between CHAID and RFM may seem minor and potentially inconsequential from a practical standpoint. However, assessing these differences in potential revenues can provide valuable context. By analysing response data using hypothetical cost and profit figures, one can gauge the effectiveness of RFM and CHAID segmentation algorithms.

Assuming a mailing cost of $1.50 and an expected $100 response, and considering the holdout sample as the remaining portion of the house file while the test sample involves a 10% test mailing, the results are insightful. The RFM approach applied to the test sample predicts a profit of $284,903 for the entire file, whereas CHAID predicts a profit of $277,488. Sending the complete file to probable recipients indicated by the two approaches would result in a profit of $274,781 for CHAID and $241,469 for RFM, showcasing CHAID's superiority by $33,313. These slight percentage variations can have a significant impact when applied to large datasets.

Study 2

The study utilizes data gathered from a non-profit organization's database, which recently sought donations from its members. The dataset comprises 99,200 individuals, randomly split into a holdout sample of 49,600 and a test group of 49,600. The solicitation yielded a 27.4% response rate, with response rates of 27.3% and 27.6% for the test group and holdout sample, respectively. Table 2 illustrates the percentage of respondents collected for 10% increments of file depth, ranging from 20% to 50%, using three segmentation strategies for both the test and holdout groups. It also presents the difference in proportions between the test and holdout samples for each segmentation method at each depth. Across all four depth levels, the proportions of the test and holdout samples for each of the three procedures are nearly identical, differing by less than one percentage point, with no statistically significant variations. These findings suggest that all three approaches accurately estimate the response rate for this specific dataset, which exhibits a high response rate.

Table 2 also displays the gain percentage for each segmentation technique across the four test depth levels and holdout samples. RFM and CHAID values remain consistent between the test and holdout samples across all four depth levels. Interestingly, when CHAID is constrained to the same independent factors as RFM, RFM appears equally accurate as CHAID in identifying potential responders in this dataset.

An unexpected discovery concerns the effectiveness of logistic regression in terms of gain. CHAID outperforms logistic regression significantly at both the 20% and 30% depth levels for the test and holdout samples. Similarly, RFM outperforms logistic regression at a depth of 30%. Consequently, logistic regression may not be as effective for this specific dataset, especially if a marketer decides to mail to a portion of the data.

Table 2 *Percent of Total Responses at Various Levels*

	Data-Mining Technique		
	RFM (%)	**CHAID (%)**	**Logistic (%)**
20% depth			
Test Sample	35.3ab	35.6a	34.7b
Hold Sample	34.2ab	35.2a	34.2b
Difference	1.1	0.4	0.5
30% depth			
Test Sample	48.9a	48.9a	45.8b
Hold Sample	47.6a	48.7a	45.8b
Difference	1.3	0.2	0.0
40% depth			
Test Sample	59.2a	58.4a	58.2a
Hold Sample	58.4a	58.2a	58.6a
Difference	0.8	0.2	0.4
50% depth			
Test Sample	70.4a	69.4a	68.4a
Hold Sample	70.1a	69.2a	68.3a
Difference	0.3	0.2	0.1

Research 2 findings indicate that when response rates are relatively high, RFM might exhibit performance akin to more sophisticated statistical methods. Hughes' RFM method can achieve gain levels of likely responders and logistic regression comparable to CHAID across all four file depth test levels.

4. Discussion

The two studies present intriguing findings regarding the effectiveness of the three segmentation methods. They depict two distinct scenarios that database marketers may encounter based on the datasets, each yielding slightly different outcomes. One scenario involves a recent solicitation mail that garnered responses from approximately 25% of the total file, while the other pertains to contributions solicited by Study 2, a non-profit organization, from their house file. In this context, RFM demonstrates comparable accuracy to CHAID and logistic regression in capturing likely responders across tested file depths (20% to 50%) and exhibits equally reliable parameters when applied to the holdout sample. Consequently, based on the results of study 2, it can be inferred that RFM effectively segments potential respondents similar to the other two methods, especially in scenarios with a high response rate.

However, research 1's characteristics present a more typical scenario for database marketers, with a response rate below 5% for a mail-order company dataset. In such prevalent direct marketing scenarios, the results suggest that RFM may not perform as effectively as CHAID, particularly when mailing to a small portion of the file (30% or less). CHAID outperforms RFM in these scenarios in terms of dependability and identifying potential responders. Notably, CHAID accumulates a greater number of responses compared to RFM at the 20% and 30% file depths, suggesting that statistical algorithms like CHAID might excel in categorizing dataset members compared to RFM's predetermined groupings.

The study underscores the importance of considering database marketing situations wherein RFM may be less effective than CHAID. While RFM may struggle in scenarios with low response rates and a need to target a small percentage of the database, it can still perform well in high-response rate scenarios or when targeting a large portion of the database. This versatility makes RFM a cost-effective and dependable alternative, particularly when basic transaction characteristics are the primary focus. However, it's essential to recognize that the conclusions drawn from these studies may not apply universally to all database marketing situations.

Furthermore, the study points out limitations in the analysis of RFM, particularly its narrow focus on historical individual behavior. This limitation may hinder a marketer's understanding of customers' broader motivations and values. To address this, it suggests incorporating relational data alongside transactional data, utilizing analytical techniques like logistic regression and CHAID that can accommodate a wider range of customer attributes. Overall, future research should explore these segmentation techniques across various contexts, considering diverse predictor-response relationships and broader customer insights beyond transactional data alone.

5. References

1. Baier M, Ruf KM, Chakraborty G. Contemporary database marketing: concepts and applications. Evanston: Racom Communications; 2002.
2. Deichmann J, Eshghi A, Haughton D, Sayek S, Teebagy N. Application of Multiple Adaptive Regression Spines (MARS) in direct response modelling. J Interact Market 2002;16:15–27 [Autumn].
3. Drozdenko RG, Drake PD. Optimal database marketing: strategy, development, and data mining. Thousand Oaks: Sage Publications; 2002.
4. Hughes AM. Strategic database marketing. New York: McGraw–Hill; 2000.
5. Kahan R. Using database marketing techniques to enhance your one-to-one marketing initiatives. J Consum Mark 1998;15:491–4.
6. Levin N, Zahavi J. Predictive modeling using segmentation. J Interact Market 2001;15:2-22 [Spring].
7. Linder R, Geier J, Kolliker M. Artificial neural networks, classification trees and regression: which method for which customer base? Database Mark Cust Strategy Manag 2004;11:344–56 [July].
8. Magidson J. Improved statistical techniques for response modeling: progression beyond regression. J Direct Mark 1988;2:6-17 [Autumn].
9. Marcus C. A practical yet meaningful approach to customer segmentation. J Consum Mark 1998;15:494–501.
10. McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. Journal of business research, 60(6), 656-662.
11. Sargeant A, McKenzie J. The lifetime value of donors: gaining insight through CHAID. Fund Rais Manag 1999;30:22–7 [March].
12. Verhoef PC, Spring PN, Hoekstra JC, Leeflang PS. December the commercial use of segmentation and predictive modelling techniques for database marketing in The Netherlands. Decis Support Syst 2002;34:471–81.
13. Yang AX. How to develop new approaches to RFM segmentation. J Target Meas Anal Mark 2004;13:50–60 [October].
14. Zahay D, Peltier J, Schultz DE, Griffin A. The role of transactional versus relational data in IMC programs: bringing customer data together. J Advert Res 2004;44:318 [March].