NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Direct Marketing Campaign Optimization:

Multi-objective Approach with RFM Model

Chekit Mehta, Corvus Lee, Paritosh Sharma, Yining Feng

Master's in Data Science, Northwestern University

MSDS 460: Decision Analytics

Dr. Nathan Bastian

August 26, 2020

Abstract

Data mining is a powerful technique to aid companies showcase different patterns and trends in their customers data and then help them drive improved customer relationships. This is the basis for customer relationship management (CRM). Organizations today are utilizing data analytics to help structure their direct marketing campaigns and implement targeted promotions for their customer base, given the vast amount of data generated by both online and offline via various purchase methods. Using data analytics, companies can implement effective strategies that can help define market segmentation, personalize promotional offers, allocate funding for marketing efficiently, and above all improve the relationship with their customers. Using historical data on customer purchases, the proposed model identifies customer segments based on the classic recency, frequency, and monetary value (RFM) model. The objective of this paper is to demonstrate the use of a goal programming approach and determine the customer segments that should be targeted to achieve profit maximization. However, there are some drawbacks for data mining tools, such as neural networks have long training times and genetic algorithms are brute computing methods. This study proposes goal programming, to extract meaningful rules, and it can effectively improve these drawbacks by gaining prioritized outputs. Then, considering different marketing model priorities, the goal programming model helps identify the profile segments most worthy to target. The implementation of such strategies is often governed by limited budgets and the ever-changing priorities and goals of marketing campaigns. This is also further dependent on a couple of constraints, like different priorities on models and budget allocations for a direct marketing campaign.

Keywords: RFM, integer programming, goal programming, direct marketing, customer lifetime value

Introduction

In today's times, CRM is all about obtaining and managing customer data that can guide them with customers' characteristics, buying habits, and buying potential. These types of data sources can be multifold like internally generated data, public databases or third party vendors who sell these types of data. Direct marketing that can utilize these models generated via widespread use of data analytics allowing them to use this customer data for fine tuning different marketing strategies with precision and accuracy. Data analytics involves strategic and quantitative analysis to improve business decision making. Important attributes of these customer data that can be utilized in direct marketing with focus on improvements can be to improve response rates, conversion rates, and campaign profitability. One particular analytical tool used frequently in direct marketing is the RFM model. This recency, frequency, monetary value (RFM) framework creates highly customized promotional campaigns to reach these customers. Hence leading to highly effective direct marketing campaigns by enabling companies to further categorize customers into homogenous segments based on their previous purchasing behavior. The resulting homogeneous segments targeted for promotional offers are created to assist with outreach activities. For example, if a given customer segment shows a low value for recency and higher values for frequency and monetary value then these customers are typically approached with a "we want you back" marketing strategy. If a given customer segment shows a low monetary value and high values for frequency and recency, a more relevant "up-selling" marketing strategy could be designed to generate additional sales revenue. The RFM model typically assumes that a company can reach all its customers, even customers with less than optimal RFM scores. In reality, this is not going to be the case as organizations operate under yearly budget constraints, and therefore such assumptions are not real. Hence it would be wise to add optimization to the RFM approach that can help allocate resources in effective distribution. This project highlights a multi-objective optimization methodology based on a goal programming (GP) approach to profit maximization for direct marketers using RFM data. One unique characteristic of this (GP) model is the inclusion of varying direct marketing objectives as well as corresponding budget constraints.

Literature Review

The RFM criteria have been used for many years as an analytical technique because, as measures of customers' prior behavior, they are key predictors of their future purchase behavior. The RFM model involves choosing customers based on when they last purchased (recency), how often they purchased (frequency), and how much they spent (monetary value) on past purchases. It is considered to be simple and generally reliable (McCarty & Hastak, 2007). Asllani & Halstead (2011) offered an integer linear programming approach, which remains applicable when companies do not have complete transaction records, using a) recency only, b) recency and frequency, and c) recency, frequency and monetary value of customer's purchases. A later research (Asllani & Halstead, 2015) used a goal programming approach, which was novel in combining marketing priorities and preferences for given customer segments while recognizing the reality of annual spending limits on direct marketing programs. Another paper (Asllani &

Lari, 2015) touched on the RFM analysis from a goal programming standpoint and discusses customer lifetime value (CLV) and linear programming for marketing campaigns. It also highlights advantages and disadvantages of RFM —– basically why segments achieved from RFM are not worthy of pursuit because they lack profitability.

Methodology

Data Pre-processing

The UCI machine learning repository: Online retail II data set (Dua & Graff, 2019) entails 8 variables (attributes) as shown in Table 1 of Appendix A, and it contains all the transactions occurring between December 1st, 2009 and December 9th, 2011 on a UK-based and registered non-store online retail website, which mainly sells unique all-occasion gifts.

In order to conduct the required RFM model-based clustering analysis, the original dataset needs to be pre-processed. The main steps and relevant tasks involved in the data preparation are as follows, and the Python code in Appendix B:

1. Select appropriate variables of interest from the given dataset. In our case the following five variables have been chosen: *Invoice*, *InvoiceDate*, *Customer ID*, *Quantity*, and *Price*.
2. Create an aggregated variable named *MonetaryValue*, by multiplying *Quantity* with *Price*, which gives the total amount of money spent per product/item in each transaction.
3. Sort out the dataset by *Customer ID* and create three essential aggregated variables *Recency*, *Frequency*, and *Monetary*. Calculate the values of these variables per customer.

RFM Models

The RFM (recency, frequency, and monetary value) framework is one of the most widely applied customer segmentation models used in direct marketing in which the probability of customers' future purchases is estimated to be a function of the recency, frequency, and monetary value of their previous transactions. Recency refers to the time of a customer's most recent purchase. The recency value for each customer was estimated by subtracting the snapshot date from the date where the transaction occurred. Frequency is defined as the number of a customer's past purchases, and the frequency value for each customer was based on the sum of the number of transactions. Monetary value for each customer is calculated as the sum of the purchase amount of all transactions.

A customer's RFM score is calculated based on the respective quartile rank of its recency, frequency, and monetary values. As shown in Table 2 of Appendix A, we have sorted customers by their days from last purchase in a descending order, with the most recent purchasers located at the bottom rank. Since customers are assigned with scores from 1~4, the top 25% of customers, who shopped long ago, receive a recency score of 1, the next 25% of customers receive a score of 2, and so on. Similarly, we then sorted customers by their frequency values from the most to least frequent purchasers, assigning the top 25% a frequency score of 4, etc. For the monetary value factor, the top 25% of customers (big spenders) will be assigned with a score of 4 and the bottom 25% of customers receive a score of 1. Finally, we ranked these customers by combining their individual recency (R), frequency (F), and monetary value (M) quartile ranks to arrive at an aggregated RFM score. This RFM score, displayed in Table 2 of

Appendix A, is simply the sum of the individual R, F, and M ranks. On the basis of the RFM score, we grouped the customers to 8 segments, with each having similar RFM rank, as shown in Table 3 of Appendix A.

In this dataset, the transactions created from June 1st, 2011 to August 31st, 2011 are reserved as the testing dataset to calculate probability of purchase and the expected revenue in the coming quarter. Over that particular period, there were 82,276 valid transactions in total. On the other hand 377,419 transactions happened from June 1st, 2010 to May 31st, 2011 are selected as the training dataset used for calculating RFM segments. The ratio of the total number of samples for training and the total number of samples for testing was set to 78%:22%.

Mathematical Formulation

The objective of the optimization models is to maximize profit from future purchases within the marketing campaign budget. All variables below can be calculated from sales transactions data and historical marketing campaign data.

Notations

$i = 1..R$ index to indicate a specific recency group

$j = 1..F$ index to indicate a specific frequency group

$k = 1..M$ index to indicate a specific monetary group

$V_{jk} =$ Expected revenue from a customer in frequency group j and monetary group k

$p_{ijk} =$ Probability of purchase for customers in recency group i, frequency group j and monetary group k

$N_{ijk} =$ Number of customers in recency group i, frequency group j and monetary group k

$C =$ Average cost to reach a customer

$B =$ Total budget of the marketing campaign

Integer Linear Programming (ILP) Formulation

In this section, we will set only one objective without any priority — to maximize the profit from future purchases. This serves as the best case scenario, as adding extra objectives and/or constraints will only have negative effects on the objective value.

Let the decision variables be:

$x_{ijk} = 1$ if the customer segment of recency i, frequency j, and monetary k is reached; 0 otherwise

The objective function is:

Maximize:

$$Z = \sum_{i=1}^{R} \sum_{j=1}^{F} \sum_{k=1}^{M} N_{ijk}(p_{ijk}V_{jk} - C)x_{ijk} \tag{1}$$

subject to:

$$\sum_{i=1}^{R} \sum_{j=1}^{F} \sum_{k=1}^{M} N_{ijk}Cx_{ijk} \leq B \tag{2}$$

with:

$$x_{ijk} = \{0, 1\} \quad i = 1..R \;\; j = 1..F \quad k = 1..M \tag{3}$$

Equation (1) is the sum of profits of all RFM segments, which the direct marketing campaign has reached ($x_{ijk} = 1$). Equation (2) makes sure the campaign costs less than the assigned budget B, with the left hand side the actual cost.

Multi-objective Goal Programming (GP) Formulation

In this formulation, the objective is to prioritize the direct marketing campaigns to “At Risk Customers (244)” and “Can’t Lose Them (144)” customer segments characterized with a recency score of 1 or 2, a frequency score of 3 or 4, or a monetary value score of 3 or 4.

Let the decision variables be:

$x_{ijk} = 1$ if the customer segment of recency i, frequency j, and monetary k is reached; 0 otherwise

The objective functions are:

Maximize:

$$Z = \sum_{i=1}^{2} \sum_{j=1}^{F} \sum_{k=1}^{M} N_{ijk}(p_{ijk}V_{jk} - C)x_{ijk} \qquad \text{(4) weight } W_1$$

$$Z = \sum_{i=1}^{R} \sum_{j=3}^{4} \sum_{k=1}^{M} N_{ijk}(p_{ijk}V_{jk} - C)x_{ijk} \qquad \text{(5) weight } W_2$$

$$Z = \sum_{i=1}^{R} \sum_{j=1}^{F} \sum_{k=3}^{4} N_{ijk}(p_{ijk}V_{jk} - C)x_{ijk} \qquad \text{(6) weight } W_3$$

subject to:

$$\sum_{i=1}^{R} \sum_{j=1}^{F} \sum_{k=1}^{M} N_{ijk}Cx_{ijk} \le B \qquad (7)$$

with:

$$x_{ijk} = \{0, 1\} \quad i = 1..R \;\; j = 1..F \quad k = 1..M \qquad (8)$$

Equation (4) is the sum of profits of the priority 1 RFM segments (with recency score 1 or 2) which the direct marketing campaign has reached ($x_{ijk} = 1$). Equation (5) is equivalent for priority 2 segments (with frequency score 3 or 4) and equation (6) for priority 3 segments (with monetary value score 3 or 4). The corresponding weights ($W_1,\ W_2,\ W_3$) shall be used in our computation experiments while setting the multiple objectives. Equation (7) makes sure the campaign costs less than the assigned budget B, with the left hand side the actual cost.

Computational Experiment and Results

To illustrate the models, we will be using the Gurobi optimizer 9.0 with Python 3.7 to solve the problems, and R 4.0 to perform Monte Carlo simulations and analyze the results. Assuming the retail company is running a promotional campaign every quarter, we will make use of the transactions in the last 3 months (i.e., testing dataset) to calculate the expected revenue and probability of purchase, and the past 12 months of data (i.e., training dataset) to calculate the RFM segments.

V_{jk} = total revenue in each FM segment (recency is deliberately ignored) / N_{ijk}

p_{ijk} = number of customers in the testing dataset / N_{ijk}

The expected revenue (V_{jk}) and probability of purchase (p_{ijk}), however is an uncertainty. In the Monte Carlo Simulation for the ILP and GP experiments, we would assume p_{ijk} is normally distributed, with standard deviation 0.05 and V_{jk} is uniformly distribution across $(0.5V_{jk},\ 1.5V_{jk})$. The total budget B is assumed to be \$7000 and average cost to reach a customer (which can be calculated from historical marketing data) is assumed to be \$3.5.

Figure 1

*Customer counts in each RFM segment. Color indicates V*p*

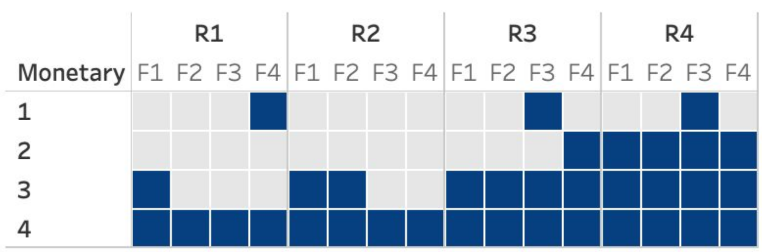
	Recency / Frequency															
	R1				R2				R3				R4			
Monetary	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4	F1	F2	F3	F4
1	359	100	27	1	193	83	35	0	146	39	14	0	69	21	7	0
2	97	188	66	8	56	167	90	26	52	133	36	20	41	78	32	4
3	22	47	83	22	29	66	174	60	19	60	162	77	13	54	150	56
4	7	5	24	21	5	11	34	82	12	19	66	224	5	15	78	486

Applying ILP to the RFM data

By employing ILP to the dataset, the customer segments colored in blue in figure 2 should be reached. The simulation shows an average total profit of $1,478,382 with 86.6% that will be greater than $1.2M. The Python code used for ILP and R code for simulation are in Appendix B.

Figure 2

Customer segments that should be reached (1: blue) or not (0: grey) base on ILP method

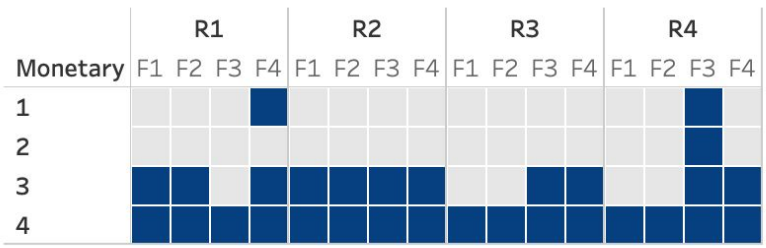


Applying Multi-objective GP to the RFM data

In this scenario we have assigned weights ($W_1 = 200, W_2 = 100, W_3 = 50$) to the first, second and third objectives respectively. This will prioritize customer segments characterized with a recency score of 1 or 2, a frequency score of 3 or 4, or a monetary value score of 3 or 4. The customer segments colored in blue in figure 3 should be reached. The simulation shows an average total profit of $1,456,657 with 84.0% that will be greater than $1.2M. The Python code used for ILP and R code for simulation are in Appendix B.

Figure 3

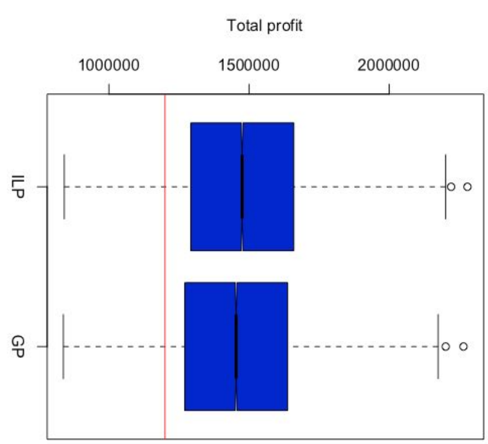
Customer segments that should be reached (1) or not (0) base on multi-objective GP method



The output shows a clear shift of focus according to the set priorities, from recency score 3 and 4 to 1 and 2. The GP model at the same time aims at optimizing the total profit, with a slight drop compared with the ILP experiment. By adjusting the objective weights, the probability of getting more than $1.2M profit will differ, reflecting the risk of the corresponding priority changes.

Figure 4

Total profits simulated using customer segments suggested by ILP and GP experiments



Discussion and Conclusions

Maximize marketing ROI (return on investment) is usually one of the top goals for any marketing organization if not for the company. As R&D investments increase and innovation drives the differentiation, there is an increasing pressure for maximizing these ROIs, and chief marketing officers (CMOs) everywhere have been forced to reduce budgets as a continuing trend over the past decade. At the same time, the direct marketing industry is currently outpacing the overall economy (DMA, 2013), representing almost 53 percent of all U.S. advertising expenditures in 2012, spending over $168 billion (accounting for 8.7 percent of GDP) and generating a ROI of over $12 for every dollar spent (Direct Marketing Association, 2012). In this paper, researchers propose an approach to formulate customer behavior and allocate marketing budgets to market segments. The top five direct marketing agencies earned over $3.5 billions in 2011, and that represented only their U.S. revenue. Thus, direct marketing continues to play an effective and growing role in the overall marketing arsenal of many organizations. In this scenario, we formulate an approach in this paper where we use available data in an integrated model as well as management judgment. This study contributes to research and practical implications, as it provides an approach to allocate marketing budgets. This research provides a practical method that marketing managers can evaluate marketing plans and make the best allocation of budgets. As CMOs are increasingly forced to achieve superior results with inferior budgets, analyzing marketing data and subsequently prioritizing marketing spending become even more crucial. Low response rates in direct marketing make budget constraints an even greater challenge for the direct response firm (e.g., 1-4 percent average response for direct mail to outbound telemarketing). Investing scarce resources on customers who are not yet willing to buy (a Type II error) is not only inefficient, but could represent a possible threat to a firm's long-term financial viability (Ferrante, 2009; Venkatesan & Kumar, 2004). The most practical advantage that this paper represents is a method where its applicability is in situations of targeting customers who haven't been in recency, but likely to make a come back purchase. The goal optimization approach used in this research achieves a balance between Type I (lack or recency of profitable customers) and Type II (possibility of higher spending by customer). It helps identify both appropriate and inappropriate RFM segments based on three core characteristics: profitability, marketing objectives, and budget constraints. By finding the most profitable customer segments (given various marketing objectives and spending limits), a GP approach applied to RFM data can provide a firm with optimal solutions to and flexibility in marketing spending decisions in a single model. Depending upon a given RFM segment's lack of recency and possible spending potential, a marketing firm can determine whether to continue targeting that segment in efforts to generate even more sales, or whether it should spend its scarce resources on alternative (i.e., more profitable) groups.

This research can therefore be used as a type of scoring model for practitioners to enable the transformation of purchasing history data, i.e., RFM data, into a useful decision model which can be applied to many marketing situations and to any imposed budget limitation. Because this research factors in budget constraints and different marketing priorities, the decision model demonstrated here has considerable long-term utility for maximizing the profitability of customer segments.

This study has limitations, but these can provide avenues for future research in the area. For example, because RFM frameworks represent historical behavior, their ability to accurately capture and predict future behavior and profit potential has been questioned (Blattberg et al., 2009; Rhee & McIntyre, 2009). While predicting any consumer behavior, using any type of model, is inherently uncertain (and this GP model is no exception), accuracy is always a potential limitation when forecasting is based on historical data. As the current model addresses only a two-year time period, and Venkatesan et al. (2007) argue that up to three years is considered an acceptable horizon for estimates in customer selection models, this may perhaps mitigate forecasting accuracy concerns. In other words, the shorter the time horizon considered, the less variation there is likely to be between past and future purchasing behavior (i.e., there is less time and opportunity for intervening exogenous variables to disrupt behavioral patterns). Managers sometimes were not sure about their estimations; especially, because of the uncertain macroeconomic situations and political issues of the country. Sensitivity analysis shows that only a few parameters are important for careful estimation; on the other hand, the other parameters may not have significant effects on results even if they are not estimated accurately. As noted by Davenport et al. (2010, p. 159), however, a company must still constantly review and manage its analytical models, be alert to “model decay,” monitor relevant external events, and keep track of all competing models.

Ideally, firms will eventually integrate additional customer data with RFM data as RFM focuses on customer purchasing behavior, not necessarily customer search behavior. With respect to future data collection, direct marketing managers should consider capturing web browsing data as well as transactional data, e.g., “X percent of customers clicking on Link Y ultimately visited Site Z and purchased Brand A.” This helps identify customers' search behaviors, choice criteria, and decision-making paths, all of which help us understand customer behavior better, and therefore predict it more accurately.

The value of any customer data is in how it is analyzed and then used to inform managers and help them make better business decisions (Franks, 2012). The GP approach used in this RFM analysis offers several advantages to direct marketers. It’s simple, easy to use, and can account for a large number of variables, constraints, and objectives.

References

Accenture. (2008) Most U.S. Companies Say Business Analytics Still Future Goal, Not Present Reality. Accenture press release, December 11, https://newsroom.accenture.com/subjects/research-surveys/most-us-companies-say-business-analytics-still-future-goal-not-present-reality.htm

Advertising Age. (2013) Retailers Embrace Data Tools for Rapid-Fire Price Changes after Big Holiday Season Push, Trend is Here to Stay. January 24, 2013, https://adage.com/article/digital/retailers-embrace-data-tools-rapid-fire-price/239366

Asllani, A., & Halstead, D. (2011). Using RFM data to optimize direct marketing campaigns: A linear programming approach. *Academy of Marketing Studies Journal*, *15*(2).

Asllani, A., & Halstead, D. (2015). A multi-objective optimization approach using the RFM model in direct marketing. *Academy of Marketing Studies Journal*, *19*(3).

Asllani, A., & Lari, A. (2015). Using goal programming to increase the efficiency of marketing campaigns. *Journal of International & Interdisciplinary Business Research*, *2*, 53–68.

Berger, P. D. and Nasr, N.I. (1998). Customer Lifetime Value: Marketing Models and Applications. Journal of Interactive Marketing. 12(Winter), 17-30.

Berry, M.J.A. and Linoff, G. (2004). Data Mining Techniques, 2ed. Indianapolis, IN: John Wiley & Sons.

Bhaskar, T., Subramanian, G., Bal, D., Moorthy, A., Saha, A. and Rajagopalan, S. (2009) An Optimization Model for Personalized Promotions in Multiplexes. Journal of Promotion Management. 15(1/2), 229-246.

Blattberg, R.C., Malthouse, E.C., andNeslin, S.A. (2009) Customer Lifetime Value: Empirical Generalizations and Some Conceptual Questions. Journal of Interactive Marketing. 23(2), 157-168.

Bolton, R.N. (1998) A Dynamic Model of the Duration of the Customer's Relationship with a Continuous Service Provider: The Role of Satisfaction. Marketing Science, 17(1), 45-65.

Brinker, S. (2012) "We are the 1% of global marketing spend. July 27, 2012, https://chiefmartec.com/2012/07/we-are-the-1-of-global-marketing-spend/

Ching, W.-K., Ng, M. K., and So, M. K. (2004) Customer Migration, Campaign Budgeting, Revenue Estimation: The Elasticity of Markov Decision Process on Customer Lifetime Value. Advanced Modeling and Optimization, 6(2), 65-80.

Davenport, T. H. and Harris, J. G. (2007) Competing on Analytics: The New Science of Winning. Boston, MA: Harvard Business School Press

Davenport, T. H. and Harris, J. G. (2008) Competing with Analytics. Boston, MA: Harvard Business School Publishing.

Davenport, T. H., Harris, J. G., and Morison, R. (2010) Analytics at Work: Smarter Decisions, Better Results. Boston, MA: Harvard Business School Press.

Direct Marketing Association (DMA) Press Release. (2012) DMA Releases Quarterly Business Review (QBR) for Q3 2012. November 20, 2012, http://www.the-dma.org/cgi/disppressrelease?article=563

Direct Marketing Association (DMA) Press Release. (2013) DMA Releases New Tower of Direct' Report; DM Driven Sales Growth Outpace Overall Economic Growth. March 14, 2013, http://www.thedma.org/cgi/dispannouncements?article=1590

Direct Marketing News. (2012) Direct Marketing News 2012 Agency Business Report. June 2012, http://www.dmnews.com/2012-agency-business-report/section/2738

Dua, D., & Graff, C. (2019). *UCI machine learning repository: Online retail II data set*. University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml/datasets/Online+Retail+II

Dwyer, R. F. (1989) Customer Lifetime Valuation to Support Marketing Decision Making. Journal of Direct Marketing. 3(11), 6-13.

Dye III, Thomas A. (2003) Direct Marketing Database Strategy. Atlanta, GA: Gulf Atlantic Consulting, Inc.

Edala, S. (2012), Big Data Analytics: Not Just for Big Business Anymore, December 28, https://www.forbes.com/sites/ciocentral/2012/12/28/big-data-analytics-not-just-for-big-business-anymore/#676bd8611a4b

Fader, P. S., Hardie, B. G. S., and Lee, K. L. (2005a) Counting Your Customers the Easy Way: An Alternative to the Pareto/NBD Model. Marketing Science. 24(Spring), 275-284.

Fader, P. S., Hardie, B. G. S., and Lee, K. L. (2005b) RFM and CLV: Using Iso-Value Curves for Customer Base Analysis. Journal of Marketing Research. 42(November), 415-430.

Franks, B. (2012) Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics. Wiley and SAS Business Series, Hoboken, NJ: John Wiley & Sons.

Ferrante, A. (2009) New DMA Response Rate Study Shows Email Still Strong for Conversion

Rates. DemandGen Report, The Scorecard for Sales & Marketing Automation, https://www.demandgenreport.com/features/solution-spotlight/new-dma-response-rate-study-shows-email-still-strong-for-conversion-rates

Forbes, T. (2007) Valuing Customers. Journal of Database Marketing & Customer Strategy Management. 15(1), 410.

Hambelton, K. (2013) Big Data's Effects on Direct Marketing. March 7, http://blog.neolane.com/direct-marketing2/big-datas-effects-direct-marketmg/

Kaplan, R.S., and Norton, D.P. (1992) The Balanced Scorecard-Measures that Drive Performance. Harvard Business Review. 70(January-February), 71-79.

Kumar, V., Ramani, G. and Bohling, T. (2004) Customer Lifetime Value Approaches and Best Practice Applications. Journal of Interactive Marketing. 18(3), 60-72.

Martinez, Juan. (2011) Marketers Ramp Up Real-Time Strategies. Direct Marketing News. December 1, 2011,

http://www.dmnews.eom/marketers-ramp-up-real-time-strategies/artide/217508/#

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. https://doi.org/10.1016/j.jbusres.2006.06.015

Reinartz, W. J. and Kumar, V. (2000) On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. Journal of Marketing. 64(October), 17-35.

Rhee, E. and McIntyre, S. (2009) How Current Targeting Can Hinder Targeting in the Future and What To Do About It. Journal of Database Marketing & Customer Strategy Management. 16(1), 15-28.

Sridharan, S., Frankland, D., and Smith, A. (2012) Use Customer Analytics to Get Personal, Analytically Driven Personalization Increases Retention and Return. Forrester Research Report, February 17, Cambridge, MA.

Sridharan, S., Frankland, D. and Smith, A. (2011) Cl Teams: Blocking and Tackling Is Not Enough, Forrester Research Report, July 19, Cambridge, MA.

Stahl, H.K., Matzler, K., and H.H. Hinterhuber. (2003) Linking Customer Lifetime Value with Shareholder Value. Industrial Marketing Management 32(4), 267-79.

Venkatesan, R. and Kumar, V. (2004) A Customer Lifetime Value Framework for Customer Selection and Resource Allocation Strategy. Journal of Marketing 6S(October), 106-25.

Venkatesan, R. Kumar, V., and Bohling, T. (2007) Optimal Customer Relationship Management Using Bayesian Decision Theory: An Application for Customer Selection. Journal of Marketing Research 44(November), 579-594.

Wong, E. (2009, March 3) Study: Direct Mail Spending Down 3 Percent, https://www.adweek.com/brand-marketing/study-direct-mail-spending-down-3-105372/

Appendix A

Table 1

Variables in the customer transaction dataset

Variable Name	Data Type	Description; typical values and meanings
Invoice	Nominal	Invoice number; a six-digit integral number uniquely assigned to each transaction
Stock code	Nominal	Product (item) code; a five-digit integral number uniquely assigned to each distinct product
Description	Nominal	Product (item) name
Quantity	Numeric	The quantity of each product (item) per transaction
Invoice Date	Numeric	The day and time when each transaction was generated
Price	Numeric	Product price per unit in sterling
Customer ID	Nominal	Customer identification number; a five-digit integral number uniquely assigned to each transaction
Country	Nominal	Delivery address country; United Kingdom

Table 2

Example of customer's RFM scores

Quartile	Recency Rank	Frequency Rank	Monetary Value Rank	RFM Rank	RFM Score
First Quartile	1	4	4	144	9
Second Quartile	2	3	3	233	8
Third Quartile	3	2	2	322	7
Fourth Quartile	4	1	1	411	6

Table 3

Example of RFM metrics for each customer segment

Customer Segment	RFM Rank	Characteristics
Best Customers	444	Highest frequency as well as monetary value with least recency
Loyal Customers	344	High frequency as well as monetary value with good recency
Potential Loyalists	434	High recency and monetary value, average frequency
Big Spenders	334	High monetary value but good recency and frequency values
At Risk Customers	244	Customers shopped less often now who used to shop a lot
Can’t Lose Them	144	Customers shopped long ago who used to shop a lot
Recent Customers	443	Customers who recently started shopping a lot but with less monetary value
Lost Cheap Customers	122	Customers shopped long ago but with less frequency and monetary value

Appendix B

Data pre-processing script written in Python 3.7

```
import pandas as pd

# Download the dataset from
# https://archive.ics.uci.edu/ml/datasets/Online+Retail+II

# Import dataset
print('Importing dataset')
df = pd.read_excel("online_retail_II.xlsx", sheet_name=None)
df_all = df["Year 2009-2010"].append(df["Year 2010-2011"], ignore_index=True)

# Make a copy to avoid modifying the original
retail = df_all.copy()

# Extract date
retail["InvoiceDate"] = pd.to_datetime(
      retail["InvoiceDate"], format="%d-%m-%Y %H:%M"
).dt.date

# Remove transactions without CustomerID and duplicate transactions
print('Removing transactions without CustomerID, then deduplicate')
retail = retail[~retail["Customer ID"].isnull()].drop_duplicates()

print('Split into testing and training dataset')
# Reserve one quarter of transactions from June to August 2011
# to calculate probability of purchase and the expected revenue
df_test = retail[
      retail["InvoiceDate"].between(
      pd.Timestamp("2011-06-01"), pd.Timestamp("2011-08-31")
      )
].copy()

# Take 1 year of transactions from June 2010 to May 2011
# to calculate RFM segments
df_train = retail[
      retail["InvoiceDate"].between(
      pd.Timestamp("2010-06-01"), pd.Timestamp("2011-05-31")
      )
].copy()

# Check the shape of training and testing dataset
print(df_test.shape, df_train.shape)

print('Calculating R, F, M values')
# Calculate revenue per transaction
df_train["MonetaryValue"] = df_train["Quantity"] * df_train["Price"]
```

```
# Calculate recency
max_date = max(df_train["InvoiceDate"])
df_train["Recency"] = max_date - df_train["InvoiceDate"]

# Prepare the Recency, Frequency, MonetaryValue columns
df_train_rfm = df_train.groupby(["Customer ID"]).agg(
      {"Recency": min, "Invoice": "count", "MonetaryValue": sum}
)
df_train_rfm.columns = ["Recency", "Frequency", "MonetaryValue"]

# create labels and assign them to quantile membership
print('Calculating R, F, M segments')
r_labels = range(4, 0, -1)
r_groups = pd.qcut(df_train_rfm.Recency, q=4, labels=r_labels)
f_labels = range(1, 5)
f_groups = pd.qcut(df_train_rfm.Frequency, q=4, labels=f_labels)
m_labels = range(1, 5)
m_groups = pd.qcut(df_train_rfm.MonetaryValue, q=4, labels=m_labels)

# make a new column for group labels
df_train_rfm["R"] = r_groups.values
df_train_rfm["F"] = f_groups.values
df_train_rfm["M"] = m_groups.values

print('Calculating expected revenue and purchase probability')
# Number of customers and expected revenue
# in each segment who have made new purchases (i.e., in df_test)
df_test["ExpectedRevenue"] = df_test["Quantity"] * df_test["Price"]
df_temp = df_train_rfm.join(
      df_test.groupby("Customer ID")["ExpectedRevenue"].sum(), how="inner"
)
df_test_rfm = df_temp.groupby(["R", "F", "M"]).agg({"ExpectedRevenue":
"count"})
df_test_rfm.columns = ["ReturningCustomerCount"]

# Number of customers in each segment
df_left = df_train_rfm.groupby(["R", "F", "M"]).agg({"R": "count"})
df_left.columns = ["CustomerCount"]

# Calculate expected revenue for each (F, M) segment in the "next" quarter
# We want customers to come back so (R) should be ignored
df_right = df_temp.groupby(["F", "M"])["ExpectedRevenue"].mean()

# Left join (RFM <- FM)
model_data = df_left.join(df_right).reset_index().set_index(["R", "F", "M"])
model_data.columns = ["CustomerCount", "ExpectedRevenue"]

# Calculate the purchase probability
model_data["Probability"] = (
      df_test_rfm["ReturningCustomerCount"] / model_data["CustomerCount"]
)
```

```
print('Exporting files')
# Output
model_data.fillna(0, inplace=True)
model_data.to_csv("model_data.csv")
df_temp.to_csv("df_temp.csv")
```

ILP script written in Python 3.7 with Gurobi Optimizer 9.0

```
import pandas as pd
import gurobipy as gp
from gurobipy import GRB


# Base data
df = pd.read_csv('./model_data.csv')

"""
      R  F  M  CustomerCount  ExpectedRevenue  Probability
0   4  1  1            69    296.609596  0.347826
1   3  1  1            146   296.609596  0.171233
2   2  1  1            193   296.609596  0.145078
3   1  1  1            359   296.609596  0.061281
4   4  1  2            41    308.125397  0.439024
"""

customer_count = {(df['R'][i], df['F'][i], df['M'][i]) :
                   df['CustomerCount'][i] for i in range(len(df))}
expected_revenue = {(df['R'][i], df['F'][i], df['M'][i]) :
                     df['ExpectedRevenue'][i] for i in range(len(df))}
purchase_prob = {(df['R'][i], df['F'][i], df['M'][i]) :
                  df['Probability'][i] for i in range(len(df))}
rfm_segment = list(customer_count.keys())

# Making up additional data
avg_cost = 3.5
budget = 7000

# Model
m = gp.Model("direct_marketing_ILP")

# Decision variables:
# if this rfm segment should be reached (1, 0)
rfm = m.addVars(rfm_segment, name='rfm', vtype=GRB.BINARY)

# Objective:
# maximize profit
m.setObjective(rfm.prod(
```

```
      {key: (purchase_prob[key] * expected_revenue[key] - avg_cost)
            * customer_count[key] for key in rfm_segment}
), GRB.MAXIMIZE)

# Constraints
# Max. budget
m.addConstr(
      (rfm.prod(customer_count) * avg_cost) <= budget, 'budget constraint'
)

# Compute optimal solution
m.optimize()


# Print solution
def print_solution():
      if m.status == GRB.OPTIMAL:
      print('Objective value:', m.objVal)
      m.printAttr(['x'])
      m.printAttr(['Sense', 'Slack', 'RHS'])
      else:
      print('No solution')


# Result
print_solution()
```

Simulation script written in R 4.0 for the ILP experiment

```
# Init
set.seed(1234)
DEBUG=FALSE

# Variables
df <- read.csv('model_data.csv')
avg_cost <- 3.5

# Optimized decision variables
target <- c(0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0,
            1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
            0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0,
            1.0, 1.0, 1.0, 0.0, 1.0, 1.0, 1.0, 1.0,
            1.0, 1.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0,
            1.0, 1.0, 0.0, 0.0, 1.0, 1.0, 1.0, 1.0,
            0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 0.0, 0.0,
            1.0, 1.0, 0.0, 0.0, 1.0, 1.0, 1.0, 1.0)

df$target <- target
```

```
# Function that simulate for 1 iteration
simulate_func <- function(df){
  total_profit <- 0

  for (i in 1:nrow(df)){
      if(DEBUG){cat('Simulate', i, 'segment\n')}
      total_profit <- total_profit + calc_profit(df, i)
  }

  return(total_profit)
}

# Function that returns profit of a customer segment
calc_profit <- function(df, n){
  prob <- rnorm(1, df$Probability[n], sd=.05)
  customer_count <- df$CustomerCount[n]
  expected_revenue <- runif(1, df$ExpectedRevenue[n]*.5,
df$ExpectedRevenue[n]*1.5)
  target <- df$target[n]
  profit <- customer_count * (expected_revenue * prob - avg_cost) * target
  return(profit)
}

# Simulate with n iterations
n <- 50000
total_profit <- list()

cat('Running simulation for', n , 'times\n')
print('------------------------------')

for (i in 1:n){
  total_profit <- c(total_profit, simulate_func(df))  # append the simulation
result to the list
}

# Analyze the result
total_profit <- as.numeric(total_profit)
total_profit.bar <- mean(total_profit)
hist(total_profit)
abline(v=total_profit.bar, col='red')
text(x=total_profit.bar+100000, y=7000, round(total_profit.bar))
abline(v=1200000, col='blue')
text(x=1300000, y=7000, sum(total_profit>1200000)/50000)
```

GP script written in Python 3.7 with Gurobi Optimizer 9.0

```
import pandas as pd
import gurobipy as gp
```

```
from gurobipy import GRB


# Base data
df = pd.read_csv('./model_data.csv')

"""
      R  F  M  CustomerCount  ExpectedRevenue  Probability
0   4  1  1             69    296.609596  0.347826
1   3  1  1             146   296.609596  0.171233
2   2  1  1             193   296.609596  0.145078
3   1  1  1             359   296.609596  0.061281
4   4  1  2             41    308.125397  0.439024
"""

customer_count = {(df['R'][i], df['F'][i], df['M'][i]) :
                   df['CustomerCount'][i] for i in range(len(df))}
expected_revenue = {(df['R'][i], df['F'][i], df['M'][i]) :
                   df['ExpectedRevenue'][i] for i in range(len(df))}
purchase_prob = {(df['R'][i], df['F'][i], df['M'][i]) :
                   df['Probability'][i] for i in range(len(df))}
rfm_segment = list(customer_count.keys())

# Making up additional data
avg_cost = 3.5
budget = 7000

# Model
m = gp.Model("direct_marketing_GP")

# Decision variables:
# if this rfm segment should be reached (1, 0)
rfm = m.addVars(rfm_segment, name='rfm', vtype=GRB.BINARY)

# Objective:
# maximize profit
m.ModelSense = GRB.MAXIMIZE

# Goal 1: Keep only R with 1 or 2 (weight: 200)
m.setObjectiveN(rfm.prod(
      {key: (purchase_prob[key] * expected_revenue[key]  - avg_cost)
      * customer_count[key] if key[0] in (1, 2) else 0
      for key in rfm_segment}),
      0,  weight=200, name='prioritize_low_r')

# Goal 2: Keep only F with 3 or 4 (weight: 100)
m.setObjectiveN(rfm.prod(
      {key: (purchase_prob[key] * expected_revenue[key] - avg_cost)
      * customer_count[key] if key[1] in (3, 4) else 0
      for key in rfm_segment}),
      1,  weight=100, name='prioritize_high_f')
```

```
# Goal 3: Keep only M with 3 or 4 (weight: 50)
m.setObjectiveN(rfm.prod(
      {key: (purchase_prob[key] * expected_revenue[key] - avg_cost)
      * customer_count[key] if key[2] in (3, 4) else 0
      for key in rfm_segment}),
      2,  weight=50, name='prioritize_high_m')

# Constraints
# Max. budget
m.addConstr(
      (rfm.prod(customer_count) * avg_cost) <= budget, 'budget constraint'
)

# Compute optimal solution
m.optimize()


#Print solution
def print_solution():
      if m.status == GRB.OPTIMAL:
      decision = m.getAttr('x')
      print('Objective value: ',
            (df['CustomerCount'] * df['ExpectedRevenue'] * df['Probability'])
            @ decision)
      m.printAttr(['x'])
      m.printAttr(['Sense', 'Slack', 'RHS'])
      else:
      print('No solution')


# Result
print_solution()
```

Simulation script written in R 4.0 for the GP experiment

```
# Init
set.seed(1234)
DEBUG=FALSE

# Variables
df <- read.csv('model_data.csv')
avg_cost <- 3.5

# Optimized decision variables
target <- c(0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
            0.0, 0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
            0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
            0.0, 0.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
```

```
            1.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0,
            1.0, 1.0, 1.0, 0.0, 1.0, 1.0, 1.0, 1.0,
            0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0,
            1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0)

df$target <- target

# Function that simulate for 1 iteration
simulate_func <- function(df){
  total_profit <- 0

  for (i in 1:nrow(df)){
      if(DEBUG){cat('Simulate', i, 'segment\n')}
      total_profit <- total_profit + calc_profit(df, i)
  }

  return(total_profit)
}

# Function that returns profit of a customer segment
calc_profit <- function(df, n){
  prob <- rnorm(1, df$Probability[n], sd=.05)
  customer_count <- df$CustomerCount[n]
  expected_revenue <- runif(1, df$ExpectedRevenue[n]*.5,
df$ExpectedRevenue[n]*1.5)
  target <- df$target[n]
  profit <- customer_count * (expected_revenue * prob - avg_cost) * target
  return(profit)
}

# Simulate with n iterations
n <- 50000
total_profit <- list()

cat('Running simulation for', n , 'times\n')
print('------------------------------')

for (i in 1:n){
  total_profit <- c(total_profit, simulate_func(df))  # append the simulation
result to the list
}

# Analyze the result
total_profit <- as.numeric(total_profit)
total_profit.bar <- mean(total_profit)
hist(total_profit)
abline(v=total_profit.bar, col='red')
text(x=total_profit.bar+100000, y=7000, round(total_profit.bar))
abline(v=1200000, col='blue')
text(x=1250000, y=7000, sum(total_profit>1200000)/50000)
```