

Explore RFM Analysis using SAS® Data Mining Procedures

Ruiwen Zhang, Cary, NC; Feng Liu, University of North Carolina at Chapel Hill, NC

ABSTRACT

The concept of RFM analysis, as it was first introduced by Bult and Wansbeek in 1995, still keeps its popularity in direct marketing, retail and database marketing because of its proven effectiveness (Blattberg et al. 2008) and easy for implementation. Incorporating RFM analysis into modern data mining techniques provides more powerful intelligence tools for market managers to predict customer behavior and make better recommendation to satisfy customers' needs (Derya Birant, 2011). In this paper, we use some examples to demonstrate the application of RFM analysis in data mining using SAS data mining procedures, for example, PROC CLUSER and PROC ARBOR. More predictive modeling techniques which are now commonly used in the marketing champagne, like neural network and support vector machine, can also integrate with RFM analysis easily.

INTRODUCTION

Recency (R), Frequency (F) and Monetary (M) describe customers' historical purchase behavior from the following aspects:

Recency: When did the customer make their last purchase?

Frequency: How often does the customer make a purchase?

Monetary: How much money does the customer spend?

RFM are three important purchase-related measures that are commonly used to identify customers who are likely to respond to promotions. RFM Analysis can help decision makers to 1) decide which customers to give particular offers based on the likelihood and find ways to increase their spending; 2) target lost customers or retain customers by giving them incentives.

With the development of modern data mining approaches, researchers consider the incorporation of RFM variables into modeling techniques, like clustering (Hosseini et al., 2010), neural network and decision tree (Olson et al., 2009), support vector machine (Zhang, 2012) or sequence of multiple data mining models (Derya Birant, 2011), etc. Its applications in data mining extend RFM analysis from customer profiling to prediction of customers' behavior and further provide better recommendations to customers based on their RFM scores as well as additional customer attributes.

The example dataset, transactional data of customer orders and used in this paper are given in table 1. This dataset is used in the RFM Analysis and also the Clustering Analysis. Another dataset which includes customer attributes is merged to the output from the Clustering Analysis, and the combined dataset is used in the Classification Analysis.

Variable	Description
CUSTMRID	Customer ID
LAST_RECORD	Purchase Dates
SALES	Sales Amount

Table 1: Transactional Data of Customer orders

RFM ANALYSIS

We provide a SAS macro function which segment the customers based on RFM score. The segmentation begins with recency, then frequency and finally monetary score. Before splitting the customers into five equal groups based on quantiles, we first sort the observations on recency, i.e. the last Purchase Date for each Customer ID. The top 20% of customers who have most recent purchases score 5 for Recency, and the second 20% of customers score 4, and so on. Given the score of Recency, we then sort and rank the observations within each of the five groups. Repeat the same procedure for Monetary given the score of Recency and Frequency. Eventually, all the customers have their RFM score by concatenating the Recency, Frequency and Monetary scores in a sequence. The macro function %RFM has following options and results.

```
%macro RFM (
```

```

/* ---Inputs---*/
InData = CUSTORDERS,
CustomerID = CUSTMRID,
Purchase_Date = LAST_RECORD,
Purchase_Amount = SALES,
Stats = SUM,
/* ---Outputs--- */
OutData = RFM_out,
/* ---Method for binning: Q = Quantile and B = Even bucket ---*/
Binning = Q,
/* ---Binning Options---*/
Recency_Bin_Num = 5,
Frequency_Bin_Num = 5,
Monetary_Bin_Num = 5,
);

```

	Customer Identifier	Number of Transactions	Date of Most Recent Transaction	Sum of Transaction Amounts	Recency Score (1=Least Recent, 5=Most Recent)	Frequency Score (1=Least Frequent, 5=Most Frequent)	Monetary Score (1=Lowest Amount, 5=Highest Amount)	RFM Score
1	000012002	1	18829	-828.8381	2	1	1	211
2	000237252	3	18437	547.3882	1	2	2	122
3	000312142	4	18231	699.0991	1	2	2	122
4	000331352	7	18387	2000.3100	1	3	3	133
5	000334302	26	19011	13509.2000	5	4	4	544
6	000390502	27	19023	4085.6308	5	4	4	544
7	000632242	1	18408	2655.6600	1	1	3	113
8	000757502	4	18641	29885.5020	2	2	5	225
9	000818952	4	18942	701.9101	3	2	2	322
10	000837502	70	19023	25173.5800	5	5	4	554
11	000846102	13	18947	3532.9200	3	3	3	333
12	000855152	15	19011	5655.2000	5	3	4	534
13	000872942	2	18891	221.5628	3	1	1	311
14	000969052	23	19011	433964.9400	5	4	5	545
15	001042352	72	19023	7058.0308	5	5	4	554
16	001200302	70	18992	1940.3000	4	5	3	453
17	001232602	3	18604	1173.3787	1	2	2	122
18	001263102	22	19011	5110.3800	5	4	4	544
19	001263552	1	18583	4030.1276	1	1	4	114
20	001290642	13	19011	1633.8800	5	3	3	533
21	001320352	1	18976	67.9400	3	1	1	311
22	001333442	3	18735	1239.7735	2	2	3	223
23	001420702	117	19023	6846514.6202	5	5	5	555
24	001465652	24	19023	1664.1700	5	4	3	543
25	001670152	7	18526	205277.7197	1	3	5	135
26	001720742	3	18711	21801.9802	2	2	4	224
27	001826502	14	18680	3876.8300	2	3	4	234
28	001935642	15	19023	426.6800	5	3	1	531

Table 3: Output of RFM Scores

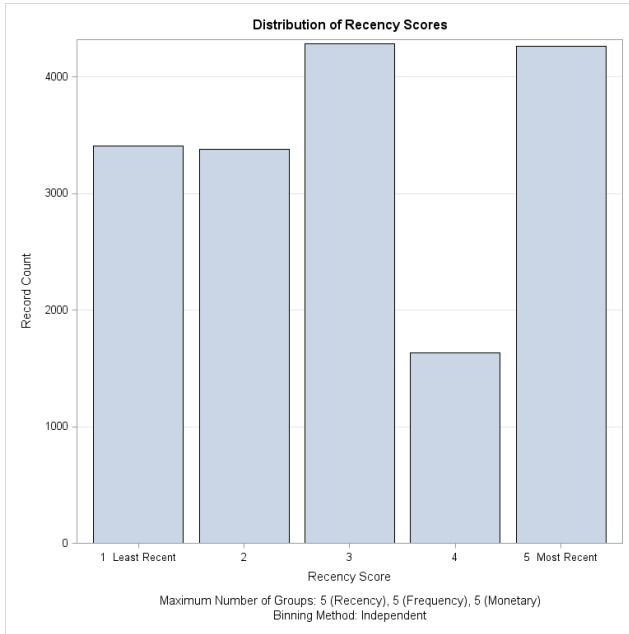


Figure 1: Distribution of Recency Scores

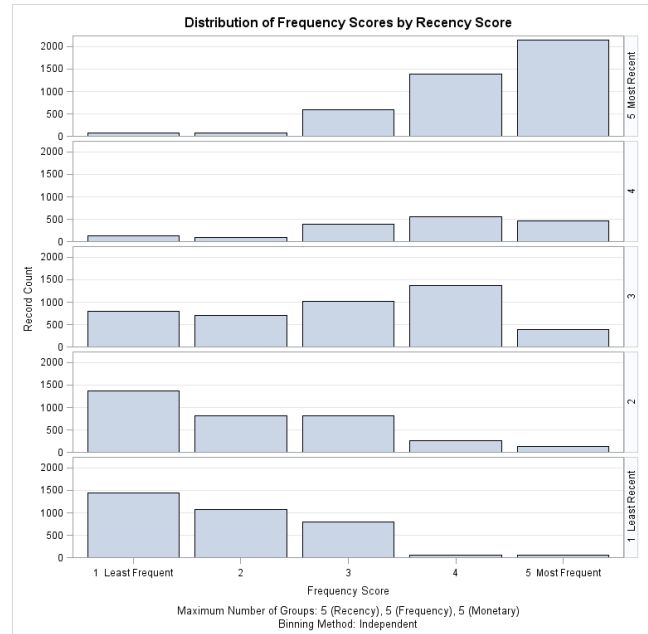


Figure 2: Distribution of Frequency Scores by Recency Score

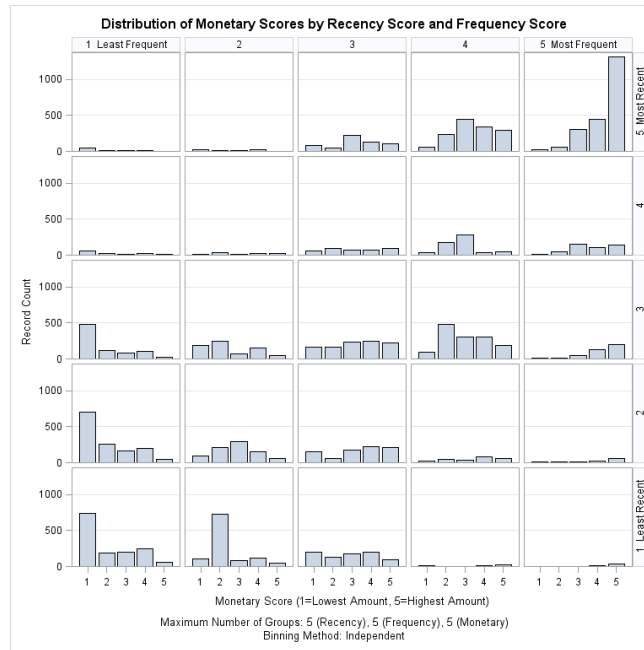


Figure 3: Distribution of Monetary Scores by Recency and Frequency Score

CUSTOMER SEGMENTATION USING RFM

After the RFM analysis, the algorithm creates total 125 (5 x 5 x 5) combinations since each attribute (R, F, M) has five scaling. Incorporating RFM into clustering analysis is one way to assign customers into appropriate cluster segments.

Customers in the same segment have similar RFM values, and so infer similar behavior of the customers within each cluster.

```
proc cluster data=rfm_out outtree=tree method=ward NOTIE k=3000 ;
    var recency_score frequency_score monetary_score;
    copy CUSTMRID rfm_score;
run;

title 'Tree Diagram using METHOD = WARD';
proc tree data = tree; id rfm_score; run;
```

Please note that the variables, Recency_score, Frequency_score and Monetary_score, are ordinal variables. All of them has values from 1 to 5, and 5 stands for the highest score. Though not same as interval variables, the ordinal variables (rank encoding) are still valid for clustering purpose.

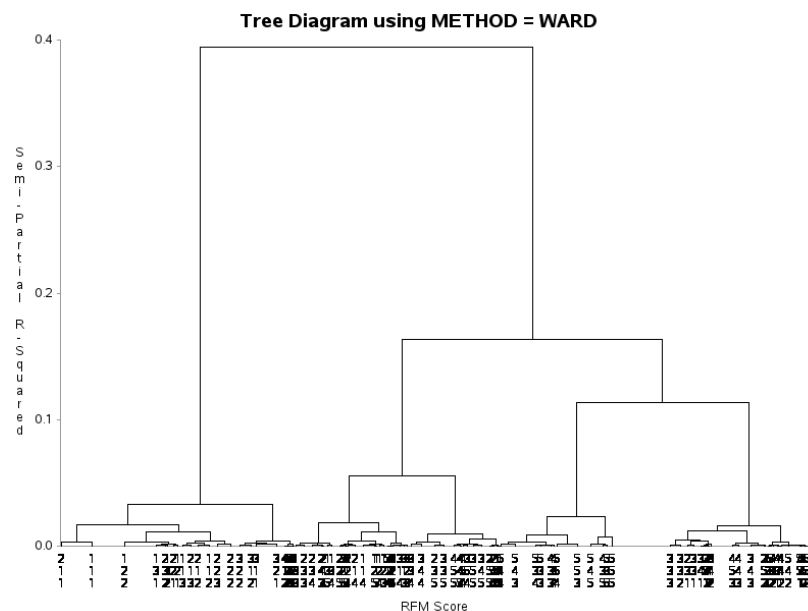


Figure 4: Cluster Analysis for Customer Segments

CLASSIFICATION FOR CUSTOMER SEGMENTS

Customers are segmented into 6 clusters based on the outputs of the clustering analysis. We then introduce the customer attributes dataset which contains predictor variables, for example, gender, income category and geometric information (country, state and city). The two datasets are merged according to CUSTOMER ID.

```
proc arbor data=clus_comb alpha = 0.2 padjust = CHAIDBEFORE DEPTH event = "6" ;
    input CITY COUNTRY STATE INCOME / level = nominal;
    input GENDER / level = nominal;
    target CLUSTER / level = nominal Criterion=PROBCHISQ;
    Performance DISK nodesize=20000 ;
    Assess NoValidata measure=MISC ;
    save MODEL=OUTTREE NODESTAT= OUTNODES TOPOLOGY= OUTTOPOLOGY;
run;
```

Target variable, the Cluster number, is specified as nominal which has 6 levels. PROC ARBOR accepts multiple input statements and each corresponds to one input level. Figure 5 shows the classification tree and each node split. The tree results can be used to classify new customers into one of the 6 segments and so target the customers with different marketing strategies.

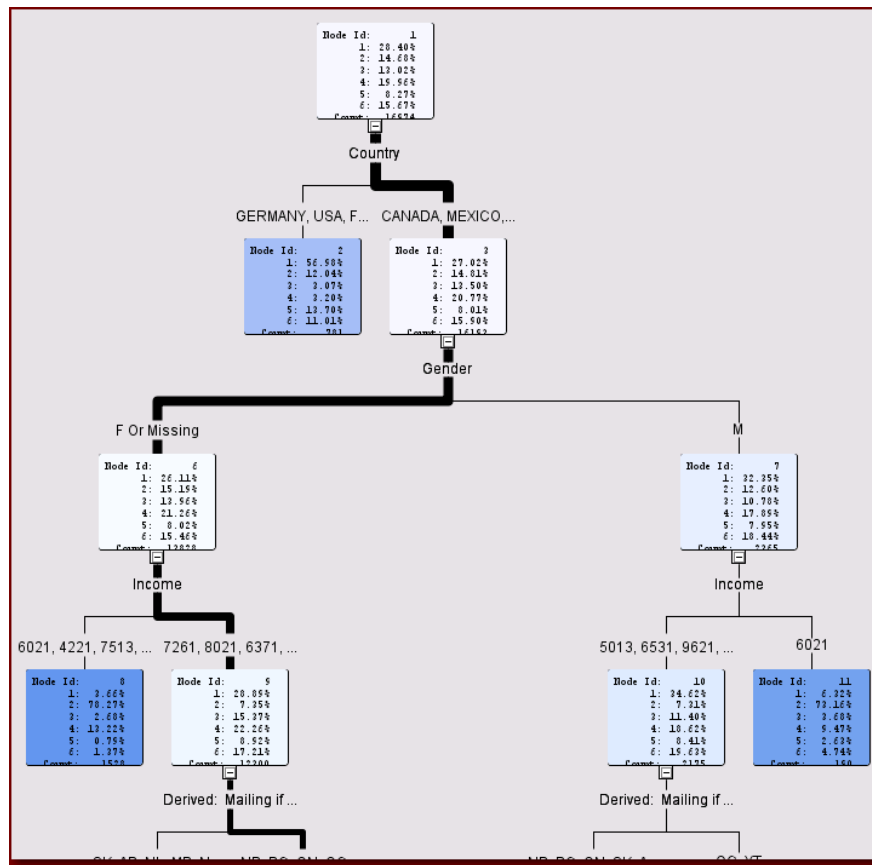


Figure 5: Decision Tree for Classification of Customer Segments

CONCLUSION

We walk through a sequence of examples starting from typical RFM analysis and then incorporate RFM in several data mining techniques and show the real applications. With less coding requirements, we can implement those applications and analyze the results using SAS data mining procedures. As future research, we are also interested to see how RFM-based variables would improve netlift modeling (Lee and et al., 2013) in terms of improving the accuracy of prediction of incremental response.

REFERENCES

- Blattberg, R.C.; Kim, B-D. and Neslin, S.A. (2008). "Database Marketing: Analyzing and Managing Customers". Chapter 12, pp. 323-337, Springer, ISBN: 978-0387725789, New York.
- Bult, J.R. and Wansbeek, T. (1995). "Optimal selection for direct mail". Marketing Science, Vol. 14, No. 4, pp. 378-394, ISSN: 0732-2399.
- Birant, Derya (2011). "Data Mining Using RFM Analysis." Knowledge-Oriented Applications in Data Mining. Prof. Kimito Funatsu (Ed.). ISBN: 978-953-307-154-1, InTech.
- Hosseini, S.M.; Maleki, A. and Gholamian, M.R. (2010). "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty". Expert Systems with Applications: An International Journal, Vol. 37 Issue 7, pp. 5259-5264

Lee, T.; Zhang, R.; Meng X. and Ryan, L. (2013). "Incremental Response Modeling Using SAS® Enterprise Miner™". SUGI proceedings, pp. 096 - 2013.

SAS/STAT(R) 9.2 User's Guide, Second Edition

Zhang, R. (2012). "A Support Vector Machine (SVM) Approach to Netlift Modeling". Joint Statistical Meetings proceedings, Section on Statistics in Marketing, pp. 3336 - 3340. American Statistical Association.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ruiwen Zhang
SAS Institute, Inc
E-mail: Ruiwen.Zhang@sas.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.