Paper 223-28

A SAS® Market Basket Analysis Macro: The "Poor Man's Recommendation Engine"

Matthew Redlon, Decision Intelligence, Inc., Eden Prairie, MN

ABSTRACT

Market basket analysis is a mathematical technique frequently used by marketing professionals to reveal affinities between individual products or product groupings. Decision Intelligence, Inc. (DII) has prepared an open source SAS market basket analysis macro and made it available for download at www.dii-online.com. The purpose of this paper is to run through data preparation, program execution and, in an effort to explain the results, one potentially valuable application. Whether for web site personalization, telemarketing or email marketing campaigns, the popularity and success of recommender systems continues to grow each year. Unfortunately, the cost of implementing such a solution is still prohibitively high for most small companies. Using the SAS market basket analysis macro, even a small company can have what we like to call the "Poor Man's Recommendation Engine". The paper describes how to filter the noise from the large quantities of association rules the macro will generate, and how to produce a look-up table to aid in making your next round of targeted recommendations for your company the most successful yet.

INTRODUCTION

Market basket analysis is a common mathematical technique used by marketing professionals to reveal affinities between individual products or product groupings. Decision Intelligence, Inc. (DII) has prepared an open source SAS Market Basket Analysis Macro and made it available for download at www.dii-online.com. The purpose of this paper is to run through data preparation, program execution and, in an effort to explain the results, one potentially valuable application.

DATA PREPARATION

DATA FORMAT

The input SAS data set must contain two columns of data. The first required column is referred to as the "Basket Dimension". This is typically a unique customer identifier. The second column is referred to as the "Analysis Unit". This is typically a product or product grouping identifier. The SAS data set should contain one row of data for each unique combination of "Basket Dimension" and "Analysis Unit" that occurred during the analysis period. A visual example can be seen in Figure 1.

	BASKET_DIMENSION	*ANALYSIS_UNIT*
1	CUSTOMER 1	PRODUCT A
2	CUSTOMER 1	PRODUCT D
3	CUSTOMER 2	PRODUCT B
4	CUSTOMER 2	PRODUCT C
5	CUSTOMER 2	PRODUCT D

Figure 1

DATA QUANTITY

The quantity of transactional data required is a function of purchase frequency, number of unique products and level of product aggregation. In our experience, a safe starting point for a market basket analysis is 12 months of transactional data at a reasonably high level of product aggregation (i.e. "t-shirts" rather than "large blue t-shirt", "small green t-shirt", etc.).

PROGRAM EXECUTION

Once a SAS data set has been created containing the transactional data using the methodology described above, several macro variables must be defined prior to program submission:

```
/*libname libref 'library
   location';*/
%let lib = work; *Library Name;
%let set = transactional_data;
   *Dataset Name;
%let basket_dimension =
unique_customer_id;
   *Basket Dimension;
%let analysis_unit = product_id;
   *Analysis Unit Identifier;
```

Define lib, set, basket_dimension, and analysis_unit where 'lib' is the library containing the transactional data set (use libname statement if necessary), 'set' is the transactional data set name, 'basket_dimension' is the unique basket dimension identifier (i.e. customer identifier) and 'analysis_unit' is the unique analysis unit identifier (i.e. product identifier).

UNDERSTANDING AND USING THE RESULTS

Whether for web site personalization, telemarketing or email marketing campaigns, the popularity and success of recommender systems continues to grow each year. Unfortunately, the cost of implementing such a solution is still prohibitively high for most small companies. The following explanation of results demonstrates the use of the market basket analysis SAS Macro to generate a very simple recommendation lookup-table.

ANALYSIS_UNIT	ANALYSIS_UNIT_FREQ	ASSOC_ANALYSIS_UNIT	ASSOC_ANALYSIS_UNIT_FREQ	FREQ_CO_OCCUR	TOT_BASKET_DIMENSIONS	SUPPORT	CONFIDENCE	EXPECTED_CONFIDENCE	LIFT
ELECTRONICS	1,192,895	TOYS	490,712	185,279	4,360,952	4.25%	15.53%	11.25%	1.38
ELECTRONICS	1,192,895	SPORTS & RECREATION	489,526	154,968	4,360,952	3.55%	12.99%	11.23%	1.16
ELECTRONICS	1,192,895	HARDWARE	556,042	175,835	4,360,952	4.03%	14.74%	12.75%	1.16
ELECTRONICS	1,192,895	ATHLETIC WEAR	353,542	111,147	4,360,952	2.55%	9.32%	8.11%	1.15
ELECTRONICS	1,192,895	COSMETICS/ ACCESSORIES	329,204	102,435	4,360,952	2.35%	8.59%	7.55%	1.14
ELECTRONICS	1,192,895	JEWELRY	540,819	158,878	4,360,952	3.64%	13.32%	12.40%	1.07
ELECTRONICS	1,192,895	APPAREL	597,537	173,235	4,360,952	3.97%	14.52%	13.70%	1.06
ELECTRONICS	1,192,895	HOME ACCESSORIES	1,087,950	281,670	4,360,952	6.46%	23.61%	24.95%	0.95
ELECTRONICS	1,192,895	HOUSEWARES	1,422,272	362,645	4,360,952	8.32%	30.40%	32.61%	0.93
ELECTRONICS	1,192,895	DOMESTICS	1,767,049	375,961	4,360,952	8.62%	31.52%	40.52%	0.78

Figure 2

Figure 2 depicts an example taken from Fingerhut customer purchases over a one year period. Fingerhut was a direct marketer of general merchandise for forty years. For each column in the output SAS data set shown in Figure 2, a brief description follows:

A) **ANALYSIS_UNIT** - This is the item or group being examined to determine if it has an affinity with the ASSOC_ANALYSIS_UNIT. Example: Our ANALYSIS_UNIT is the "Electronics" product grouping. Thus, we would like to find out what ASSOC_ANALYSIS_UNIT we should recommend to a customer who has purchased from the "Electronics" product grouping.

B) **ANALYSIS_UNIT_FREQ** - This is the number of unique customers who purchased from the ANALYSIS_UNIT. Example: 1,192,895 unique customers purchased from the "Electronics" product grouping.

C) **ASSOC_ANALYSIS_UNIT** - This is the item or group being compared in a particular affinity test with the ANALYSIS_UNIT. Example: Our ASSOC_ANALYSIS_UNIT is the "Toys" product grouping. Thus, we would like to know if we should recommend the "Toys" product grouping to someone who has purchased from the "Electronics" product grouping, relative to all other product groupings.

D) **ASSOC_ANALYSIS_UNIT_FREQ** - This is the number of unique customers who purchased from the ASSOC_ANALYSIS_UNIT. Example: 490,712 unique customers purchased from the "Toys" product grouping.

E) **FREQ_CO_OCCUR** - The number of unique customers who purchased from both the ANALYSIS_UNIT and the ASSOC_ANALYSIS_UNIT. Example: 185,279 unique customers purchased from both the "Electronics" product grouping and the "Toys" product grouping.

F) **TOT_BASKET_DIMENSIONS** - The number of unique customers who purchased from any product grouping during the 12 month time frame being investigated. Example: 4,360,952 unique customers purchased from a product grouping during the 12 months being investigated.

G) **SUPPORT** – The percentage of unique customers who purchased the ANALYSIS_UNIT and the

ASSOC_ANALYSIS_UNIT out of all unique customers who purchased during the 12 month time frame being investigated. Example: Of the 4,360,952 unique customers who purchased from a product grouping during the 12 month time frame being investigated, 185,279 purchased from both the "Electronics" product grouping and the "Toys" product grouping. This means that 4.25% of unique customers purchased from both the "Toys" and "Electronics" product grouping during the 12 month analysis period.

H) **CONFIDENCE** - The percentage of times the ANALYSIS_UNIT and the ASSOC_ANALYSIS_UNIT are purchased by the same unique customer out of all unique customers who purchase the ASSOC_ANALYSIS_UNIT. Example: Of the 1,192,895 unique customers who purchased from the "Electronics" product grouping, 185,279 also purchased from the "Toys" product grouping. This means that 15.53% of customers who purchased from the "Electronics" product grouping also purchased from the "Toys" product grouping.

I) **EXPECTED_CONFIDENCE** - The percentage of unique customers who purchased from the ASSOC_ANALYSIS_UNIT out of all unique customers who purchased during the 12 month time frame being investigated. Example: Of the 4,360,952 unique customers who purchased from a product family during the 12 month time frame being investigated, 490,712 purchased from the "Toys" grouping. This means that we would expect 11.25% of customers to have purchased from the "Toys" product grouping.

J) **LIFT** - A ratio of the confidence to the expected confidence. Values greater than 1 indicate the ANALYSIS_UNIT and the ASSOC_ANALYSIS_UNIT occur together more frequently than we would expect, while values less than 1 mean the ANALYSIS_UNIT and the ASSOC_ANALYSIS_UNIT occur together less frequently than we would expect. Example: Given a confidence of 15.53% and an expected confidence of 11.25%, we compute the lift to be: .1553/.1125 or 1.38. We expected 11.25% of customers who purchased from the "Toys" product grouping to also purchase from the "Electronics" product grouping, when in actuality 15.53% did. Lift represents the 38% increase in observed co-occurrences over expected co-occurrences.

LIMITATIONS AND CONSIDERATIONS

Obviously, an analysis at the highest level of product aggregation is of limited value. Which "Toy" should we recommend to the person who purchases from the "Electronics" product grouping? The problem you will face is that as the number of unique products to be compared increases, the length of time required by the macro to compute all possible affinities increases greatly. In its current form, the SAS macro is rendered hopelessly inefficient by data sets containing more than a few hundred unique product groupings.

The macro currently computes only one level affinities, for example, "Electronics"➔"Toys". A natural extension of the algorithm would be to calculate multi-level affinities, for example, "Electronics" and "Sports & Recreation"➔"Athletic Wear". The inclusion of multi-level affinities does, however, make the process of acting on the findings using a look-up table more difficult.

Although all of the quantities in the above examples have adequate levels of support and confidence, market basket analysis at finer levels of granularity will frequently produce rules which are either insignificant or obvious. Rules with very low Support should be excluded due to the low frequency of occurrence, which renders the rules inapplicable to the greater customer population. Extremely high or low Confidence levels, typically although not always associated with low Support levels, can lead to artificially inflated Lift values.

The bottom line is that using the results of the market basket analysis macro requires an in depth understanding of the product level data and a commitment to sifting through the potentially voluminous results to find actionable rules.

GENERATING THE RECOMMENDATION LOOK-UP TABLE

Once all insignificant and/or obvious rules have been filtered from the resulting data, the top n rules can be placed into a look-up table in the following format:

ANALYSIS UNIT	1st CHOICE	2nd CHOICE	3rd CHOICE	4th CHOICE
ELECT-RONICS	TOYS	SPORTS & REC-REATION	HARD-WARE	ATHLETIC WEAR

Figure 3

One alternative for cost effective storage and retrieval of association rules is to generate a Microsoft Access database which stores data in a format similar to Figure 3. A Microsoft Access form can then be created allowing users to input an ANALYSIS_UNIT and receive the top n rules for a particular product or product grouping. This

application would be particularly useful for telemarketing representatives capitalizing on cross-sell opportunities.

CONCLUSION

There are several marketing and merchandising applications for market basket data. A few examples from the direct marketing and retail industries are described here.

- Within a catalog, customers can be directed from one product to a strongly-associated product.
- In servicing customers on the phone, as a customer is purchasing a product, strongly-associated products can be offered.
- In paginating catalogs or in placing products in retail stores, strongly-associated products can be displayed together as a unit, or as complementary offerings.
- Products can be bundled together, and perhaps sold at a discount when the bundle is purchased (versus the individual products).
- Given past purchase history of a particular product, emails can be developed to recommend strongly-associated products.
- In general emails, perhaps to newer customers, products *not* strongly associated may be offered up (in order to demonstrate the variety of products available).
- In pricing products, margins for strongly-associated products could be adjusted. For example, if product 2 is always purchased with product 1, product 1 could be positioned with a lower margin and product 2 with a higher margin.

With the SAS macro described here, you can easily produce recommendations for these applications and more.

ACKNOWLEDGEMENTS

The author wishes to acknowledge Jeff Gunderson for his help in preparing and testing the code and process described in this paper.

CONTACT INFORMATION

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

Matthew Redlon
Decision Intelligence, Inc.
19105 Homestead Circle
Eden Prairie, MN 55346
Work Phone: 612-325-9385
Fax: 952-474-7394
Email: contact@dii-online.com
Web: www.dii-online.com