Paper 1728-2018

Retail Product Bundling – A new approach

Bruno Nogueira Carlos, Youman Mind Over Data

ABSTRACT

Affinity analysis is referred to as Market Basket Analysis in retail and e-commerce outlets application. It determines how often items are purchased together and the best possible groupings of products which are regularly bought by the customers in order to understand their behavior and preferences. This analysis allows to design effective marketing campaign and is appropriate for several kinds of decisions, such as cross-selling products, bundle of products, product placement and guide websites and loyalty program designs.

Product bundling strategy can help either to sell two or more products or services together, or to optimize marketing content, such as product catalogues, including social media and internet advertising.

The output of SAS® Enterprise Miner™ Market Basket Node generates vast numbers of rules which allow us to analyze discovered patterns that are often used as business guidelines for marketing strategies. However, this is not enough to identify the exact decreasing sequential co-occurrence of the set of items, that embodies the bundle. To overcome this subject, a new approach, implemented as a new SAS® Enterprise Miner™ node, is proposed to identify a bundle, simplifying the selection of items and adding richer information to the marketing initiatives.

INTRODUCTION

Product bundling is a process of combining two or more items into one bundled solution. In the context of marketing strategy, this process is used for packaging together several products or services. However, in some contexts of marketing, the order of sequence of the items that co-occur together in a bundle can bring valuable information regarding consumers' shopping patterns. This type of analysis can be used to guide the placement of items in leaflets, coupons, store layout, product catalogues, including social media and internet advertising or some other profitable action.

To address this challenge and provide details of how items are bundled, a new approach, implemented as a new SAS® Enterprise Miner™ node, called Bundles, is proposed to identify the bundles with the optimal sequence of the items. It gives rich information to the marketing initiatives, including details of the bundling process. The node was implemented as a SAS® Enterprise Miner™ extension node, which provides a mechanism to develop custom solutions and it is indistinguishable from any other node.

As a part of a strategy to make a marketing campaign more productive and cost-effective, the results from the Bundles Node can be used to identify either the items to be offered to a customer or the potential customers to buy an item. This avoids offering something that the consumer does not have profile to accept. This personalized marketing action is recurrent in companies which recognize their customers, for instance, in e-commerce, telecommunications, streaming services and retail with loyalty card program. It is very valuable to be used in recommendation engines and targeted marketing campaign.

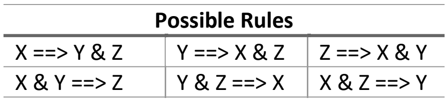
To achieve that, the Next Best Offer Node was developed, also using the SAS® Enterprise Miner™ extension node. It applies the results from the Bundles Node in a new transactional database. This process scores each customer with the most relevant items that are meaningful to each of them based on their historic behavior, ordered by those they more likely want. A classification is done to indicate if the item is an opportunity to be suggested as a “new offer” or as a “re-buy”.

This paper is intended for business analysts and marketing teams who are looking to optimize marketing content. It was developed with retail in mind but can be extended to other industries (streaming service, automotive, insurance, telecommunications, etc.).

MOTIVATION

The Market Basket Analysis, which is a popular technique available in the SAS® Enterprise Miner™ found in both Association and Market Basket Node. The technique recognizes items that are regularly bought by the customers in order to understand their behavior and preferences and also identifies the possible groupings of products (bundles) available in form of rules. A rule, according with SAS® Enterprise Miner™ 14.3: Reference Help (chapter 28) is written using the form A ==> B, where A is called the antecedent and B is called the consequent. Both, A and B, do not have items in common and each, A and B, may include one or more distinct items.

If the order of sequence of the items that are put together is not relevant, any of the rules that lead to the same bundle can be used to represent it. This allows, for instance, to group products into a single-price bundle. To illustrate that, suppose there are three items, X, Y and Z. They all together may be part of an identical bundle with size 3 (bundle containing 3 distinct items), which can be found as a different rule up to six times in the output of the Market Basket Analysis. The Figure 1 shows all those possible rules that can be found. The rules to be available on the output are subject to the setting of the node's properties.



Possible Rules		
X ==> Y & Z	Y ==> X & Z	Z ==> X & Y
X & Y ==> Z	Y & Z ==> X	X & Z ==> Y

Figure 1: Possible rules for the same bundle found in the output for the items X, Y and Z

However, the order of the items that co-occur together in a bundle is relevant for many of the marketing activities, once it brings valuable information regarding consumers' shopping patterns. In this case, it is essential to find the set of items that best represents how they are put together.

Note that the structure of the rule generated by the Market Basket Analysis relates the antecedent and the consequent of the rules, which means it does not take in account the order of the items inside both the antecedent and the consequent. For instance, the rule "Z ==> X & Y" has the consequent with the items X and Y, regardless of their order. Because of that, there no guarantee if any of the rules put together represent the best order of the items.

PROPOSED APPROACH

The Bundles Node is proposed to identify the recommended bundle with the optimal sequence, returning much less rules than the traditional approaches. It performs bundle mining over transaction data in conjunction with item taxonomy in case it is available. The transaction data contains sales transaction records detailed with items bought by customers. The frequency of the items is irrelevant, only the presence of the items matters.

This node uses the information from the transaction data to seek for expressive bundles, or set of items, with the optimal sequence by identifying the best order in which items occur together, based on several constraints including support and confidence. According with SAS® Enterprise Miner™ 14.3: Reference (chapter 28) support and confidence, are defined as:

- Support — Quantifies the frequency of transactions that contains both item P and Q as count or percentage. For the percentage, the support for the rule P ==> Q can be expressed mathematically as the following ratio:

$$\frac{Transactions\ that\ contain\ both\ item\ P\ and\ Q}{All\ transactions}$$

- Confidence — Given the rule P ==> Q, the confidence for the rule is the conditional percentage that a transaction contains item Q, given that the transaction already contains item P. Confidence for the rule P ==> Q can be expressed mathematically as the ratio:

$$\frac{Transactions\ that\ contain\ both\ item\ P\ and\ Q}{Transactions\ that\ contain\ item\ P}$$

To explain this method, suppose the items P and Q have sold 800 and 500 respectively, and 400 together. Thus, once the minimum transaction frequency (support) is set to 200 and the minimum confidence is set to 50%, one of the following rules may be the classified as a recommended: P ==> Q or

Q ==> P. It happens because they have the same support 400 which is higher than or equal to the minimum required (400) and because their confidence is also higher than or equal to the minimum required: confidence of the rule P ==> Q is 50% (400/800) and the confidence of the rule Q ==> P is 80% (400/500). Therefore, the rule Q ==> P is recommended to represent the bundle with size 2. The recommended rule is the one where the second item has the highest confidence related to the first, i.e., 80% of those who bought the first item, Q, also bought the second item, P.

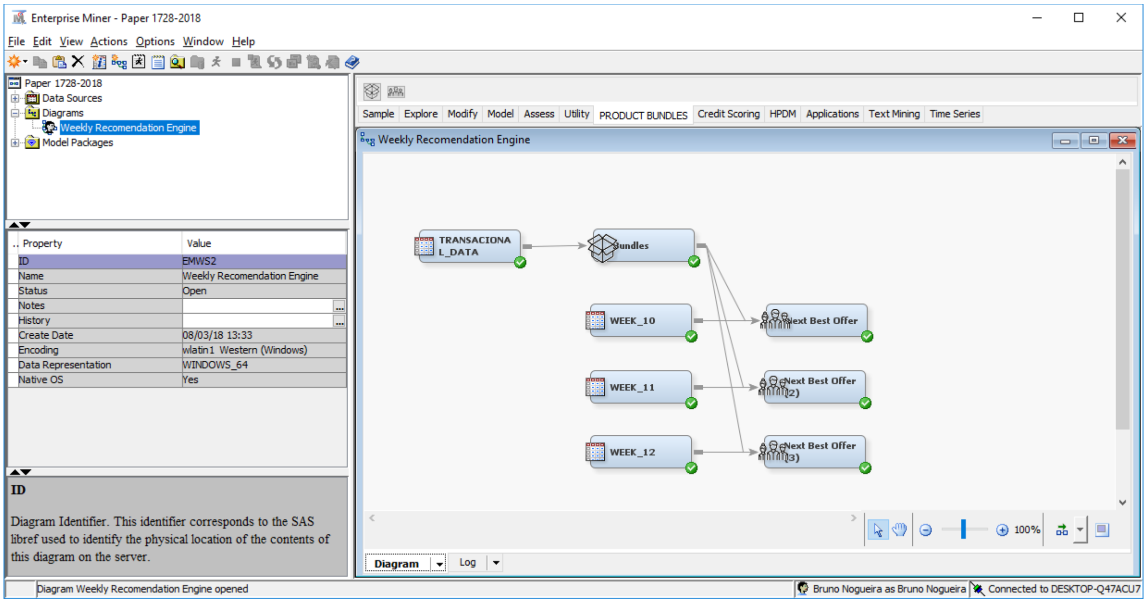
The description above refers to 2 items. It may be expanded to 3. In this case, the given bundle with size 2 embodies another item with the highest confidence related to it, classifying it as a recommended rule. So, the current bundle now has a size 3 and it refers to a bundle with a specific order of items with highest confidence along its bundling. If none of the items generates a bundle with size 3, the bundle remains with size 2. Once the bundle with size 3 is defined, another item may be combined growing the size of the bundle to 4. This process continues until the constraint of maximum items is reached or until none of the remaining items generates a bundle with one more item.

The bundles found give insights to the patterns and behavior of all customers. For further marketing actions, these insights can be applied to a new transactional data to provide personalized recommendations to each customer. As result, each item present in any bundle has its most likely customers identified. It supports the marketing team either to target the right customers in their marketing actions (targeted mailing) or to promote items to a specific customer (recommendation engine).

To attain that, the Next Best Offer (NBO) Node was developed. It allows to classify potential customers for each item according to their probability to be either a new buyer or a re-buyer. If there is no detailed information of transactions by customer, the Next Best Offer Node can still be used for general assessments.

To explain this method, for a bundle with size 2, if a customer has in its transaction history only the first item in the bundle, the customer becomes a potential buyer for the second item (new offer). If it has both, it is a potential buyer for the second item (re-buy). This explanation extends for any bundle size.

The Display 1 depicts a diagram named “Weekly Recommendation Engine” which contains an example of the application of both nodes on daily activities. These nodes are on the “Product Bundles” tab of the SAS® Enterprise Miner™ tools bar. Notice the Bundles Node was trained with the “Transactional Data”, returning recommended bundles and then it was applied in conjunction with a new transaction data, for instance, “Week 12”, to the Next Best Offer Node.



Enterprise Miner - Paper 1728-2018

Property	Value
ID	EMWS2
Name	Weekly Recomendation Engine
Status	Open
Notes	
History	
Create Date	08/03/18 13:33
Encoding	wlatin1 Western (Windows)
Data Representation	WINDOWS_64
Native OS	Yes

ID

Diagram Identifier. This identifier corresponds to the SAS libref used to identify the physical location of the contents of this diagram on the server.

Diagram Weekly Recomendation Engine opened

Bruno Nogueira as Bruno Nogueira

Connected to DESKTOP-Q47ACU7

Display 1: SAS® Enterprise Miner™ screen

The following section details both nodes regarding their output, their usage and properties.

BUNDLES NODE

As in any other SAS® Enterprise Miner™ node, the results window of the Bundles Node can be found by right-clicking the node and selecting Results from the pop-up menu.

Select View from the main menu to view besides the default Properties, SAS Results, Scoring, Table and Plot, the following results:

- Bundle Output — Contains outputs about bundles to assist the marketing team for optimize marketing content:
  - Rules Table — Gives several measures for items and bundles during its bundling.
  - Product Bundle Profile — Gives descriptive statistics about items present in bundles.
  - Rules Table per Item — Lists, by each item, all the recommended bundles where it is present.
- Hierarchy Output — Contains insights on items based on their taxonomy to improve the strategy of a marketing campaign:
  - Hierarchy Driver — The output lists ranked items by their popularity for each taxonomy.
  - Hierarchy Table — Lists, by each item, its parent Name and taxonomy level.
- Descriptive Statistics Output — Contains statistics regarding the items and the basket size:
  - Basket Profile — Presents the transaction count per basket size.
  - Item Statistics — Presents statistics about the items contained within the basket sizes.
  - Basket Profile Plot — Presents a plot of the basket sizes by transaction count (%).

The next sections describe each of these results with examples.

Bundle Output

Three outputs are given by the Bundles Node to assist the marketing team for optimize marketing content:

Rules Table

Gives several measures for items and bundle its bundling. Its columns are described as follows:

1. Rule ID: ID which identifies a bundle with a specific order of items.
2. Bundle Size: Number of items inside a bundle.
3. Columns with the prefix “Item”: depict the name of the item.
4. Value Description: Describes the measure recorded on the columns with the prefix “Value Item”. By default, these measures are support, support (%), confidence and, when it applies, the level of taxonomy. It may have further customized extra information when it is set and mapped (e.g. contribution margin, production cost and sell price).
5. Columns with the prefix “Value Item”: Shows the values of the measure indicated by the column “Value Description”. The first value (Value Item 1) refers to a bundle with one single item which is described in the column “Item 1”. The second value (Value Item 2) refers to a bundle with only two items which is described, respectively, in the columns “Item 1” and “Item 2”, and so on.
6. Recommended: indicates whether a rule is recommended.

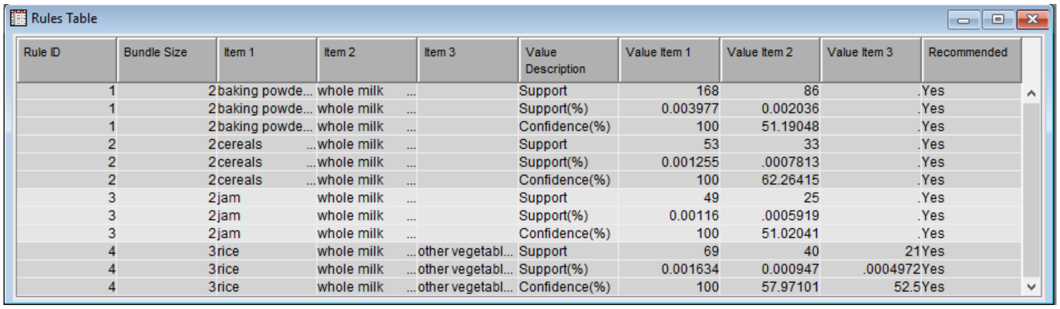
The Display 2 depicts an output example of the Rules Table. Notice that in the Rule ID number 3, there are up to 2 items in the bundle (see ‘bundle size’). The first item ‘jam’ has a support of 49 (see both ‘Value Description’ and ´Value Item 1’), which means there are 49 transactions with this item. When the second item ‘whole milk’ is bundled, the support is now 25 (see ´Value Item 2’), which means there are 25 transactions containing both items. The support (%) follow the same understanding and shows the proportion of the items bundled in all transactions. The confidence (%) of ´Value Item 2’ shows the proportion of transactions with the ‘jam’ that remains after the ‘whole milk’ is bundled. To be clear, of all ‘jam’ transactions, 51% (25/49) included also ‘whole milk’.

In the Rule ID number 4, there up to 3 items in the bundle. The first item ‘rice’ has a support of 69, which means there are 69 transactions with this item. When the second item ‘whole milk’ is bundled the support is now 40, which means there are 40 transactions containing both items. Once the third item ‘other vegetables’ is bundled the support becomes 21, which means there are 21 transactions containing those three items. The support (%) follow the same understanding and shows the proportion of the items bundled in all transactions.

The confidence (%) of ‘Value Item 2’ shows the proportion of transactions with the first item that remains after the second item is bundled. Of all ‘rice’ transactions, 57,97% (40/69) included also ‘whole milk’. The confidence (%) of ‘Value Item 3’ shows the proportion of transactions with the first and second items that also include the third item: 52.50% (21/40).

When taxonomy is present, an extra line is added for each Rule ID with Value Description of ‘Hierarchy Level’ and the columns with the prefix “Value Item” shows the taxonomy level of respective item.

If any extra information is set and mapped, additional lines for each rule are inserted. The Value Description will show the label of the extra information, defined during the mapping. The columns with the prefix ‘Value Item’ are a cumulative sum of the extra information value. To be clear, the ‘Value Item 1’ will refer to item 1 only, the value item 2 will refer to item1 + item2, and so on. If an item is not found on the extra information, the sum returns a missing value from that point.



Rule ID	Bundle Size	Item 1	Item 2	Item 3	Value Description	Value Item 1	Value Item 2	Value Item 3	Recommended
1	2	baking powde...	whole milk ...		Support	168	86	.	Yes
1	2	baking powde...	whole milk ...		Support(%)	0.003977	0.002036	.	Yes
1	2	baking powde...	whole milk ...		Confidence(%)	100	51.19048	.	Yes
2	2	cereals ...	whole milk ...		Support	53	33	.	Yes
2	2	cereals ...	whole milk ...		Support(%)	0.001255	.0007813	.	Yes
2	2	cereals ...	whole milk ...		Confidence(%)	100	62.26415	.	Yes
3	2	jam	whole milk ...		Support	49	25	.	Yes
3	2	jam	whole milk ...		Support(%)	0.00116	.0005919	.	Yes
3	2	jam	whole milk ...		Confidence(%)	100	51.02041	.	Yes
4	3	rice	whole milk ...	other vegetabl...	Support	69	40	21	Yes
4	3	rice	whole milk ...	other vegetabl...	Support(%)	0.001634	0.000947	.0004972	Yes
4	3	rice	whole milk ...	other vegetabl...	Confidence(%)	100	57.97101	52.5	Yes

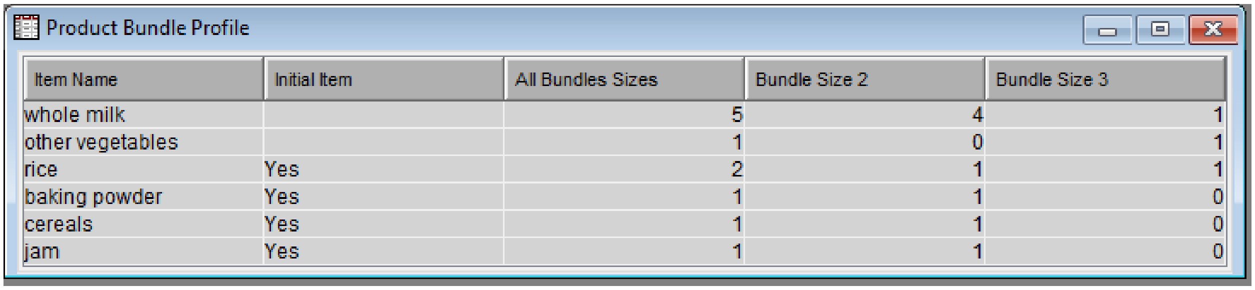
Display 2: Rules Table output example

Product Bundle Profile

Gives descriptive statistics about items present in bundles. Its columns are described as follows:

1. Item Name: Depict the name of the item.
2. Taxonomy Level: Indicates the hierarchy of the item. This column appears when the taxonomy data is set and mapped.
3. Initial Item: indicates whether an item lead to other items, which means there is at least one bundle derived from that item. Otherwise, the item does not lead to any other and should not be used as a primary selling item in a marketing campaign.
4. All Bundles Sizes: displays how many times each item belongs to different bundles. It sums up the values of the columns with the prefix “Bundles Size”.
5. Columns with the prefix “Bundle Size”: depict how many times the item belongs to specific bundle size. For instance, when an item has the value 7 recorded on column “Bundle Size 2”, means it belongs to 7 distinct bundles of size 2.

The Display 3 depicts an output example of the Product Bundle Profile. The item ‘whole milk’ does not lead to other items (see ‘initial item’), even though it is present in 5 distinct bundles: 4 with two items and 1 with three items. The item ‘rice’ leads to other items, which means it is the first item of at least one bundle.



Product Bundle Profile

Item Name	Initial Item	All Bundles Sizes	Bundle Size 2	Bundle Size 3
whole milk		5	4	1
other vegetables		1	0	1
rice	Yes	2	1	1
baking powder	Yes	1	1	0
cereals	Yes	1	1	0
jam	Yes	1	1	0

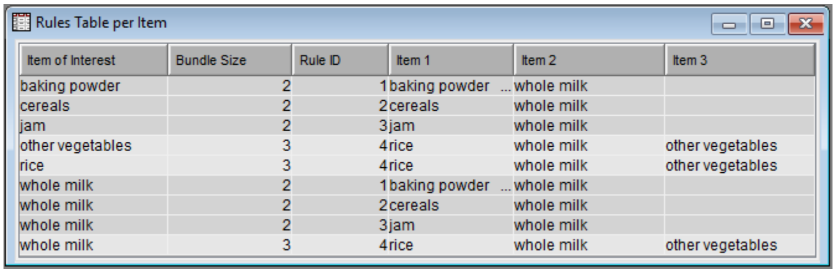
Display 3: Product Bundle Profile output example

Rules Table per Item

Lists, by each item, all the recommended bundles where it is present. Its columns are described as follows:

1. Item of Interest: Refers to an item included in the bundle.
2. Bundle Size: Number of items inside a bundle.
3. Rule ID: ID which identifies a bundle with a specific order of items.
4. Columns with the prefix “Item”: depict the name of item.

The Display 4 depicts an output example of the Rules Table per Item. In this output all items are ordered alphabetically along with the bundles in which they are included. For example, ‘whole milk’ is present in rules ID 1, 2, 3 and 4. In each rule, ‘whole milk’ is bundled with different products. Notice that rule ID 4 is also shown for items of interest ´rice´ and ‘other vegetables’.



Rules Table per Item

Item of Interest	Bundle Size	Rule ID	Item 1	Item 2	Item 3
baking powder	2	1	baking powder ...	whole milk	
cereals	2	2	cereals	whole milk	
jam	2	3	jam	whole milk	
other vegetables	3	4	rice	whole milk	other vegetables
rice	3	4	rice	whole milk	other vegetables
whole milk	2	1	baking powder ...	whole milk	
whole milk	2	2	cereals	whole milk	
whole milk	2	3	jam	whole milk	
whole milk	3	4	rice	whole milk	other vegetables

Display 4: Rules Table per Item output example

Hierarchy Output

One strategy to improve a design of a marketing campaign and catch the consumers’ attention for a specific product from a family of products, is to display or emphasis the most popular characteristic it has. Figure 2 illustrates a hypothetic example of the taxonomy for a generic soda marketed with three packings with the same flavor. The most popular packing is the can, with 68% of all transactions, followed by the one litter plastic bottle with 17% and 15% for small glass bottle. Hence, the can should be used as a primary packing to design the marketing campaign.

Notice that the packaging feature is one example of many others which can consist of flavors, fragrances, colors, and so on.

Figure 2: Hierarchy Driver example

The Bundles Node shows an output which allows an analyst to obtain those insights. This output requires that the taxonomy data be set and mapped. The parameter of the ‘maximum item’ specifies the number of items to be shown in the hierarchy driver output. For example, if 2 is set, the output will detail the first and the second top items, grouping all others as one new item called '_OTHERS_'.

Practical Application: The marketing team may consider reaching a target audience according to ongoing strategies and based on channels, weekdays, specific stores, stores' location and so forth. To accomplish that and obtain consistent results, they must filter an appropriate transaction dataset and rerun the Bundles Node.

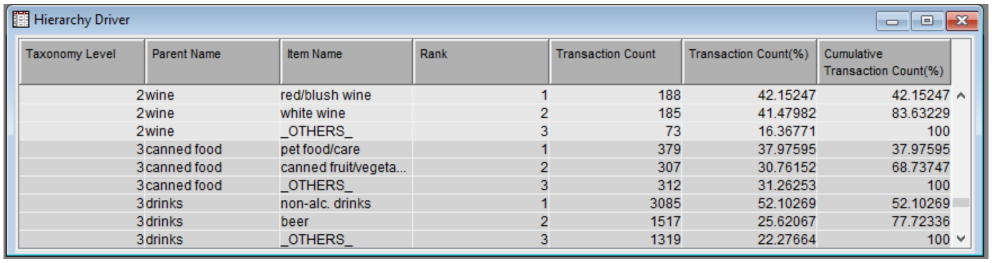
In addition to the Hierarchy driver output, also a hierarchy table is generated, detailing all taxonomy levels.

Hierarchy Driver

The output lists ranked items by their popularity for each taxonomy and, in case of a tie, they receive the same rank. The results appear as follows:

1. Taxonomy Level: Indicates the hierarchy of the item.
2. Parent Name: Contains the parent name of the item.
3. Item Name: Depict the name of the item.
4. Rank: Indicates the rank of the item based on its popularity for its hierarchy.
5. Transaction count: Number of transactions with the item.
6. Transaction count (%): Percentage of transaction count over the total of its parent item.
7. Cumulative Transaction count (%): The cumulative value of the transaction count (%).

The Display 5 depicts an output example of the Hierarchy Driver. In this example, the maximum item parameter was set to 2, hence for each taxonomy there are 2 distinct items and when necessary an item named _OTHERS_ is output grouping the remaining items. For the item wine, the two most popular are red wine with 188 transactions and white wine with 185 transactions, both account for 83,6% of all wine transactions (see 'Cumulative Transaction count (%)').



Hierarchy Driver

Taxonomy Level	Parent Name	Item Name	Rank	Transaction Count	Transaction Count(%)	Cumulative Transaction Count(%)
2	wine	red/blush wine	1	188	42.15247	42.15247
2	wine	white wine	2	185	41.47982	83.63229
2	wine	_OTHERS_	3	73	16.36771	100
3	canned food	pet food/care	1	379	37.97595	37.97595
3	canned food	canned fruit/vegeta...	2	307	30.76152	68.73747
3	canned food	_OTHERS_	3	312	31.26253	100
3	drinks	non-alc. drinks	1	3085	52.10269	52.10269
3	drinks	beer	2	1517	25.62067	77.72336
3	drinks	_OTHERS_	3	1319	22.27664	100

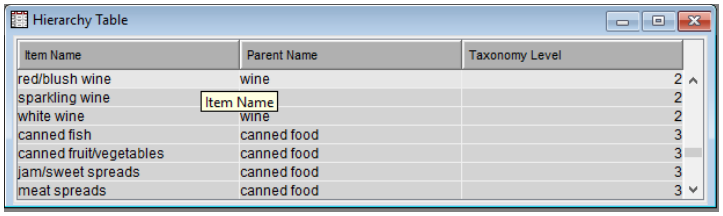
Display 5: Hierarchy Driver output example

Hierarchy Table

Lists, by each item, its parent Name and taxonomy level. Its columns are described as follows:

1. Item Name: Depict the name of the item.
2. Parent Name: Contains the parent name of the item.
3. Taxonomy Level: Indicates the hierarchy of the item.

In the display 6 'wine', on taxonomy level 2, is parent of three items: 'red/blush wine', 'sparkling wine' and 'white wine'.



Item Name	Parent Name	Taxonomy Level
red/blush wine	wine	2
sparkling wine		2
white wine	wine	2
canned fish	canned food	3
canned fruit/vegetables	canned food	3
jam/sweet spreads	canned food	3
meat spreads	canned food	3

Display 6: Hierarchy Table output example

Descriptive Statistics Output

To support the business team in decision making, some descriptive statistics related to the items and the basket sizes, i.e., a grouping of all distinct items by transaction, are generated as a means for data exploration. These metrics are merely a selection from a variety of statistics used in many companies. Other metrics can be further implemented to fit better the strategic decisions.

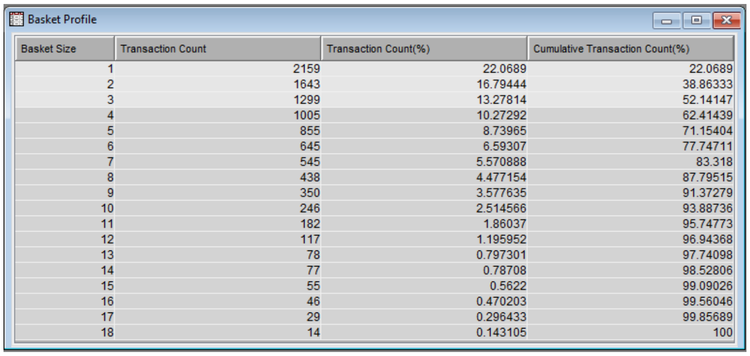
The statistics regarding the items and the basket size are in the output “Item Statistics” and “Basket Profile”. Both are contained within the results from the Bundles Node.

Basket Profile

Presents the transaction count per basket size limited to the Maximum Size property associated with Bundles Node. Its columns are described as follows:

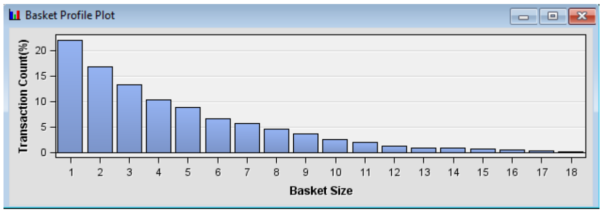
1. Taxonomy Level: Indicates the hierarchy of the items considered in the basket. This column appears when the taxonomy data is set and mapped. By default, the minimum taxonomy level is set to 1. But higher levels can be set. The results display the minimum taxonomy and all higher taxonomy levels found.
2. Basket Size: Indicates the number of distinct items.
3. Transaction count: Number of transactions.
4. Transaction count (%): Percentage of transaction count. When the taxonomy level is considered, the percent is based on the total of transaction by taxonomy level.
5. Cumulative Transaction count (%): The cumulative value of the transaction count (%).

The Display 7 depicts an output example of the Basket Profile. In this example, there are 2159 transactions with one distinct item, which represents 22.07% of all transactions in the transactional data set. Notice that more than 50% of the transactions are up to 3 distinct items. A plot is generated to depict the distribution of the basket sizes (see Display 8).



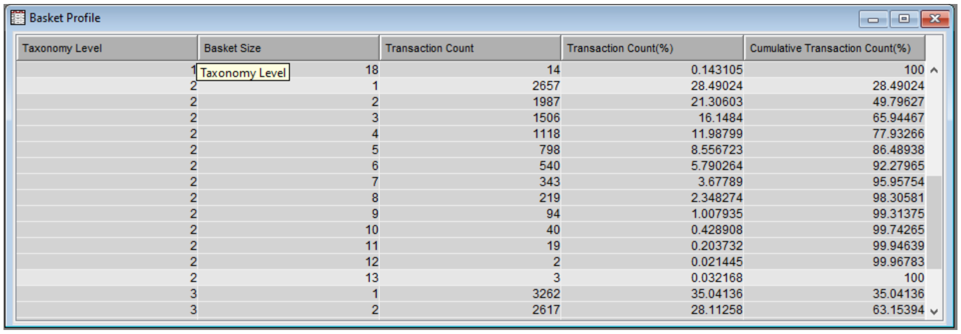
Basket Size	Transaction Count	Transaction Count(%)	Cumulative Transaction Count(%)
1	2159	22.0689	22.0689
2	1643	16.79444	38.86333
3	1299	13.27814	52.14147
4	1005	10.27292	62.41439
5	855	8.73965	71.15404
6	645	6.59307	77.74711
7	545	5.570888	83.318
8	438	4.477154	87.79515
9	350	3.577635	91.37279
10	246	2.514566	93.88736
11	182	1.86037	95.74773
12	117	1.195952	96.94368
13	78	0.797301	97.74098
14	77	0.78708	98.52806
15	55	0.5622	99.09026
16	46	0.470203	99.56046
17	29	0.296433	99.85689
18	14	0.143105	100

Display 7: Basket Profile output example without taxonomy



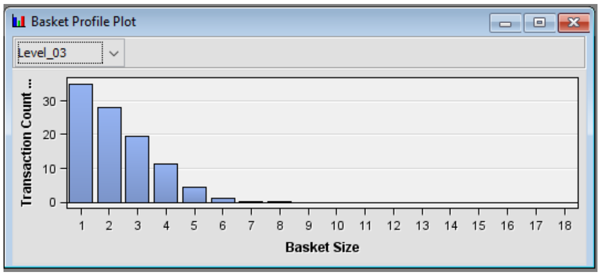
Display 8: Basket Profile Plot example without taxonomy

The Display 9 depicts another output example of the Basket Profile. In this example, the hierarchy of the items is considered in the basket. In addition to generation basket statistics for level 1, all other taxonomy levels statistics are also included in the output, bounded by the 'Minimum Taxonomy Level' property associated with Bundles Node. In this case the cumulative transaction count is made at each level. For taxonomy level 1, there are 14 transactions with 18 distinct items, if taxonomy level 2 is considered there are 2657 transactions with one distinct item. Note that are 3 transactions with 13 distinct level 2 items. A plot is generated to depict the distribution of the basket sizes by level. The level selection is on the top left corner (see Display 10).



Taxonomy Level	Basket Size	Transaction Count	Transaction Count(%)	Cumulative Transaction Count(%)
1	18	14	0.143105	100
2	1	2657	28.49024	28.49024
2	2	1987	21.30603	49.79627
2	3	1506	16.1484	65.94467
2	4	1118	11.98799	77.93266
2	5	798	8.556723	86.48938
2	6	540	5.790264	92.27965
2	7	343	3.67789	95.95754
2	8	219	2.348274	98.30581
2	9	94	1.007935	99.31375
2	10	40	0.428908	99.74265
2	11	19	0.203732	99.94639
2	12	2	0.021445	99.96783
2	13	3	0.032168	100
3	1	3262	35.04136	35.04136
3	2	2617	28.11258	63.15394

Display 9: Basket Profile output example with taxonomy level



Display 10: Basket Profile Plot example with taxonomy level

Item Statistics

Presents statistics about the items contained within the basket sizes. Its columns are described as follows:

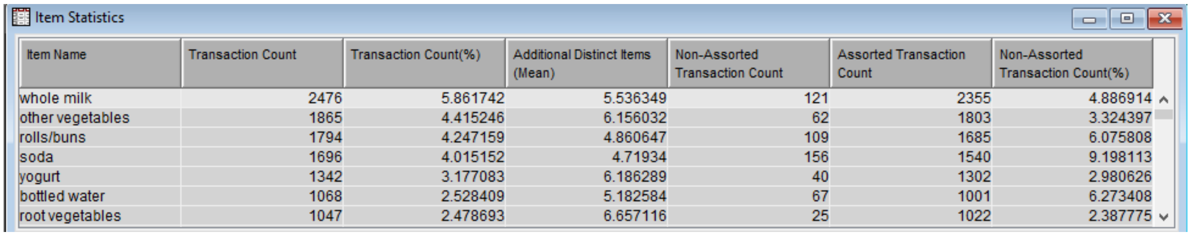
1. Item Name: Depict the name of the item.
2. Taxonomy Level: Indicates the hierarchy of the items considered in the basket. This column appears when the taxonomy data is set and mapped. By default, the minimum taxonomy level is 1. But higher levels can be set. The results display the minimum taxonomy and all higher taxonomy levels found.
3. Transaction count: Number of transactions that contain the item.
4. Transaction count (%): Percentage of transaction count. When the taxonomy level is considered, the

percent is based on the total of transactions by taxonomy level.

5. Additional Distinct Items (Mean): Average quantity of additional items that is grouped with the item.
6. Non-Assorted Transaction Count: Number of transactions in which the item is the only item in the basket.
7. Assorted Transaction Count: Number of transactions in which the item is mixed with others distinct items in the basket.
8. Non-Assorted Transaction Count (%): Percentage of Non-Assorted Transaction Count.

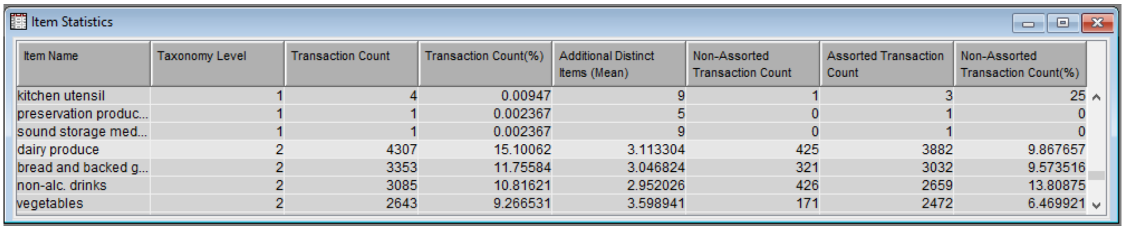
The Display 11 depicts another output example of the Item Statistics. In this example, there are 2476 transactions that have ‘whole milk’, which means 5.86% of all transactions have this item. In average, the ‘whole milk’ was grouped with 5.5 other items (see ‘additional distinct item (mean)’). In 4.89% of those 2476 transactions, ‘whole milk’ is the unique item in the transaction (see ‘non-assorted transaction count (%)’). This percentage means 121 transactions of 2476 (see ‘non-assorted transaction count’).

When taxonomy is present, all the statistics are also extended to other taxonomy levels. The Display 12 depicts another output example of the Item Statistics, there is information not only for ‘whole milk’, but also for its next level taxonomy - ‘dairy produce’. These items account for 4307 (or 15.10%) of total transactions, and in average are grouped with 3.11 other items.


Item Statistics

Item Name	Transaction Count	Transaction Count(%)	Additional Distinct Items (Mean)	Non-Assorted Transaction Count	Assorted Transaction Count	Non-Assorted Transaction Count(%)
whole milk	2476	5.861742	5.536349	121	2355	4.886914
other vegetables	1865	4.415246	6.156032	62	1803	3.324397
rolls/buns	1794	4.247159	4.860647	109	1685	6.075808
soda	1696	4.015152	4.71934	156	1540	9.198113
yogurt	1342	3.177083	6.186289	40	1302	2.980626
bottled water	1068	2.528409	5.182584	67	1001	6.273408
root vegetables	1047	2.478693	6.657116	25	1022	2.387775

Display 11: Item Statistics without taxonomy level


Item Statistics

Item Name	Taxonomy Level	Transaction Count	Transaction Count(%)	Additional Distinct Items (Mean)	Non-Assorted Transaction Count	Assorted Transaction Count	Non-Assorted Transaction Count(%)
kitchen utensil	1	4	0.00947	9	1	3	25
preservation produc...	1	1	0.002367	5	0	1	0
sound storage med...	1	1	0.002367	9	0	1	0
dairy produce	2	4307	15.10062	3.113304	425	3882	9.867657
bread and backed g...	2	3353	11.75584	3.046824	321	3032	9.573516
non-alc. drinks	2	3085	10.81621	2.952026	426	2659	13.80875
vegetables	2	2643	9.266531	3.598941	171	2472	6.469921

Display 12: Item Statistics with taxonomy level

Using the Bundles Node

The Bundles Node requires an input data source with role of transaction that contains one ID variable and one target variable. The ID variable is used to group the target into transactions (baskets). The Target variable is a nominal variable that contains the item information.

As with the Market Basket Node, a single data set with the hierarchy of items can be specified by setting the Dimension Data Set property. See the Bundles Node Properties for further details. Also, the following are not supported: Multiple parents and rugged hierarchy. According with SAS, multiple parents are ignored except the last one and if a parent exits for an item, it must immediately appear in the next level.

This node also allows to input up to 3 extra information values for each item. To do so, a single data set containing the item and a respective value (e.g. contribution margin, production cost or sell price) can be specified by setting any of the following properties: “Custom Information 1”, “Custom Information 2” or “Custom Information 3”.

For further details of how to use this node, see the section of Bundles Node Properties.

Bundles Node Properties

General Properties

The following general properties are associated with the Bundles Node:

- Node ID — The Node ID property displays the ID that SAS® Enterprise Miner™ assigns to a node in a process flow diagram. Node IDs are important when a process flow diagram contains two or more nodes of the same type. The first Bundles Node added to a diagram will have a Node ID of PRDBUNDLE. The second Bundles Node added to a diagram will have a Node ID of PRDBUNDLE2, and so on.
- Imported Data — The Imported Data property provides access to the Imported Data — Product Bundle window. The Imported Data — Product Bundle window contains a list of the ports that provide data sources to the Bundles Node. Select the … button to the right of the Imported Data property to open a table of the imported data.

  If data exists for an imported data source, you can select the row in the imported data table and click one of the following buttons:

  - Browse to open a window where you can browse the data set.
  - Explore to open the Explore window, where you can sample and plot the data.
  - Properties to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contain summary information (metadata) about the table and variables.
- Exported Data — The Exported Data property provides access to the Exported Data - Product Bundle window. The Exported Data - Product Bundle window contains a list of the output data ports that the Bundles Node creates data for when it runs. Select the … button to the right of the Exported Data property to open a table that lists the exported data sets.

  If data exists for an imported data source, you can select the row in the imported data table and click one of the following buttons:

  - Browse to open a window where you can browse the data set.
  - Explore to open the Explore window, where you can sample and plot the data.
  - Properties to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contain summary information (metadata) about the table and variables.
- Notes — Select the … button to the right of the Notes property to open a window that you can use to store notes of interest, such as data or configuration information.

Initial Settings Properties

- Variables — Select the … button to open the Variables — Product Bundle table, which enables you to view the columns metadata, or open an Explore window to view a variable's sampling information, observation values, or a plot of variable distribution. You can specify Use and Report variable values. The Name, Role, and Level values for a variable are displayed as read-only properties.

  The following buttons and check boxes provide additional options to view and modify variable metadata:

  - Apply — Changes metadata based on the values supplied in the drop-down menus, check box, and selector field.
  - Reset — Changes metadata back to its state before you click Apply.
  - Label — Adds a column for a label for each variable.

  - Mining — Adds columns for the Order, Lower Limit, Upper Limit, Creator, Comment, and Format Type for each variable.
  - Basic — Adds columns for the Type, Format, Informat, and Length of each variable.
  - Statistics — Adds statistics metadata for each variable.
  - Explore — Opens an Explore window that enables you to view a variable's sampling information, observation values, or a plot of variable distribution.
- Normalize — Use to specify whether to normalize class variables or not. This setting will result in target variable normalization where character string values are left-justified and uppercase. The default is no normalization.

Initial Settings Properties: Hierarchy

- Dimension Data Set — Use to specify the data source that defines the hierarchy of items. Select the … button to open the Select a SAS Table window.

  The data set must contain the following variables:

  - a nominal variable taking the child role
  - a nominal variable taking the parent role

  Note 1: Each distinct item that appears in the input (transactional) data set must have a parent in the lowest level of the hierarchy.

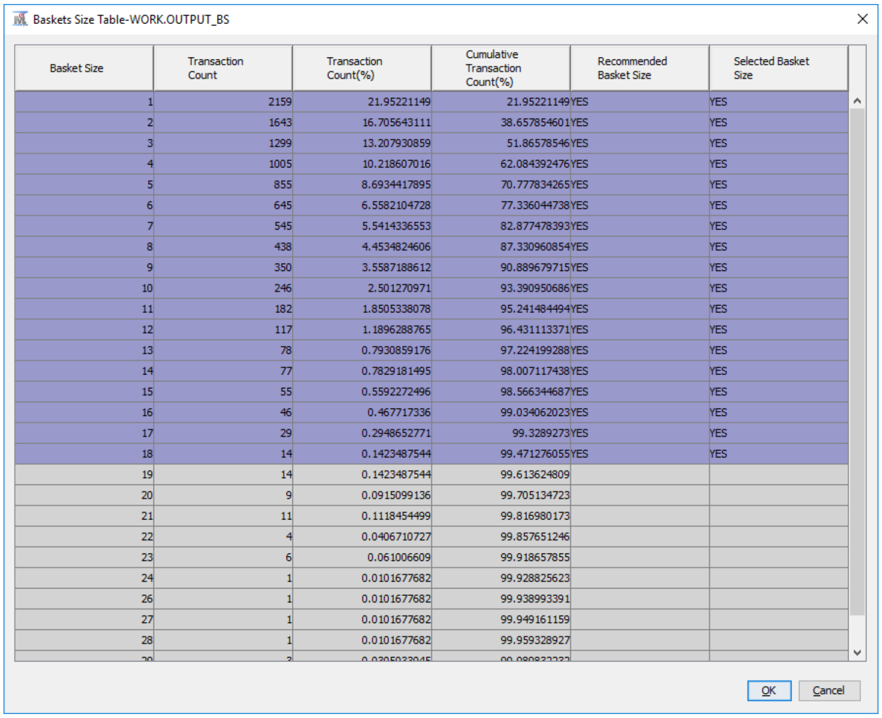
  Note 2: If no hierarchy is specified, the Bundles Node will perform simple association analysis with the input data without the hierarchy.
- Mapping — Use to open an editor that enables you to specify the parent and children variables that define the hierarchy. Select the … button to open the Mapping window.
- Hierarchy Table — Lists parent product and child defined in taxonomy database and their taxonomy levels. This table is generated after the Bundles Node run for the first time with a taxonomy data set and mapped. Use this table as guide to select the Minimum Taxonomy Level property.
- Minimum Taxonomy Level — Specifies the minimum taxonomy level to be consider valid in the training data set. The default is the lowest level: 1.

Initial Settings Properties: Basket Size Options

- Size Selection — Specifies how to select the maximum size of the basket to be considered valid in the training data set. A basket size is a grouping of all distinct target items by ID. The default setting excludes outliers that are beyond upper and lower far fences – far outliers. The User Defined setting allows to define the maximum size property.
- Maximum Size — Specifies the maximum size of the basket to be consider valid in the training data set. A basket size is a grouping of all distinct items by ID.
- Baskets Size Table — Lists basket sizes found in the current transactional dataset. This table is generated after the Bundles Node run for the first time. Use this table as guide to select the recommended basket size
  - Basket Size: Indicates the number of distinct items.
  - Transaction count: Number of transactions.
  - Transaction count (%): Percentage of transaction count. When the taxonomy level is considered, the percent is based on Minimum Taxonomy Level.
  - Cumulative Transaction count (%): The cumulative value of the transaction count (%).
  - Recommended Basket Size: It is an automatic selection made to flag and exclude far outlier

basket sizes from the analysis. Neither the outlier baskets nor its items will be included in any output

- Selected Basket Size: This selects the Maximum Size property.



Basket Size	Transaction Count	Transaction Count(%)	Cumulative Transaction Count(%)	Recommended Basket Size	Selected Basket Size
1	2159	21.95221149	21.95221149	YES	YES
2	1643	16.705643111	38.657854601	YES	YES
3	1299	13.207930859	51.86578546	YES	YES
4	1005	10.218607016	62.084392476	YES	YES
5	855	8.6934417895	70.777834265	YES	YES
6	645	6.5582104728	77.336044738	YES	YES
7	545	5.5414336553	82.877478393	YES	YES
8	438	4.4534824606	87.330960854	YES	YES
9	350	3.5587188612	90.889679715	YES	YES
10	246	2.501270971	93.390950686	YES	YES
11	182	1.8505338078	95.241484494	YES	YES
12	117	1.1896288765	96.431113371	YES	YES
13	78	0.7930859176	97.224199288	YES	YES
14	77	0.7829181495	98.007117438	YES	YES
15	55	0.5592272496	98.566344687	YES	YES
16	46	0.467717336	99.034062023	YES	YES
17	29	0.2948652771	99.3289273	YES	YES
18	14	0.1423487544	99.471276055	YES	YES
19	14	0.1423487544	99.613624809		
20	9	0.0915099136	99.705134723		
21	11	0.1118454499	99.816980173		
22	4	0.0406710727	99.857651246		
23	6	0.061006609	99.918657855		
24	1	0.0101677682	99.928825623		
26	1	0.0101677682	99.938993391		
27	1	0.0101677682	99.949161159		
28	1	0.0101677682	99.959328927		
29	3	0.0305033045	99.989832232		

Display 13: Baskets Size Table

Train Properties: Constraints

- Maximum Items — This option determines the maximum size of the item set to be considered in a bundle. Maximum allowed=50. The default is 3.

- Support Type — Specifies the type of support used for the analysis. This can either be the minimum percentage or minimum count of support to a bundle be valid:

  - Count — Use to express minimum transaction frequency as count.

  - Percent — Use to express minimum transaction frequency as percentage. This is the default. Maximum Size — Specifies the maximum size of the basket to be consider valid in the training data set. A basket size is a grouping of all distinct items by ID.

- Support Count — Use an integer value to specify the minimum transaction frequency to support. The frequency is expressed as count. This option is valid when Support Type property is set to Count. The default is 5.

- Support Percentage — Use to specify the minimum transaction frequency to support. The frequency is expressed as percentage. This option is valid when Support Type property is set to Percent. The default is 2.0%.

- Minimum Confidence Level — Use to specify the minimum confidence level that is required to group an item in a bundle. Specify an integer number between 0 and 100. The default is 50.0%.

- Recommendation — If this option is set to NO then all rules that lead to the same bundle are displayed. The default is YES, which means only one of those rules, classified as a recommended, is displayed.

Train Properties: Hierarchy Driver

- Maximum Items — Specifies the number of items to be shown in the hierarchy driver output. For example, if 2 is set, the output will detail the first and the second top items, grouping all others as one new item called '_OTHERS_'.

Train Properties: Extra Information

- Custom Information Data Set — Use to specify the data source that defines the extra information of the items. Such information can be any defined by the analyst needs. For instance, sell price, contribution margin and so on. Select the … button to open the Select a SAS Table window.

  The data set must contain the following variables for each of the custom information:

  - a nominal variable taking the item role
  - a numerical variable taking the extra information role

- Mapping — Use to open an editor that enables you to rename the label for the respective custom information and specify its item and extra information variables. Select the … button to open the Mapping window.

Status Properties

The following status properties are associated with this node:

- Create Time — displays the time at which the node was created.
- Run ID — displays the identifier of the node run. A new identifier is created every time the node runs.
- Last Error — displays the error message from the last run.
- Last Status — displays the last reported status of the node.
- Last Run Time — displays the time at which the node was last run.
- Run Duration — displays the length of time of the last node run.
- Grid Host — displays the grid server that was used during the node run.
- User-Added Node — specifies if the node was created by a user as a SAS® Enterprise Miner™ extension node.

NEXT BEST OFFER (NBO) NODE

As in any other SAS® Enterprise Miner™ node, the results window of the Next Best Offer Node can be found by right-clicking the node and selecting Results from the pop-up menu.

Select View from the main menu to view besides the default Properties, SAS Results, Scoring, Table and Plot, the following results:

- Scored Output — Two outputs are given by the Next Best Offer Node:
  - Scored Data — Lists the most likely buyers of each product and identifies the appropriate type of offer (new-offer or re-buy) to propose to each customer.
  - Scored Items Profile — Profiles the potential of customers by type of offer for each item.

The next sections describe each of these results with examples.

Scored Output

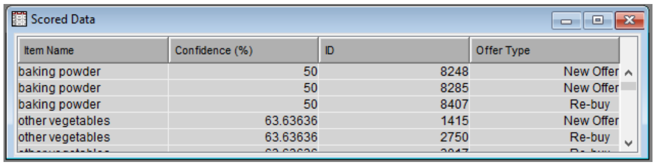
Two outputs are given by the Next Best Offer Node. One details items and potential customers and the other profiles the potential of customers by type of offer for each item:

Scored Data

List the most likely buyers of each product and identifies the appropriate type of offer (new-offer or re-buy) to propose to each customer. It is relevant by itself when there is a customer ID associated to the transactional dataset with the role of score, otherwise it must be disregarded as it is only used as a bridge to generate the report Scored Items Profile. Its columns are described as follows:

1. Item Name: Depict the name of item.
2. Taxonomy Level: Indicates the hierarchy of the item. This column appears when the taxonomy data is set and mapped on the Bundles Node.
3. Confidence (%): It is the probability of the customer buying the item given his transaction history and rules generated in the bundle output.
4. ID: This variable receives the same name of the variable with the role ID defined in the transactional data with role of score. This ID usually refers to the customer ID, if it is present. If not, it refers to the transaction ID.
5. Offer type: Indicates if the ID is eligible for a new offer of the item or for a re-buy.

The Display 14 depicts an output example of the Scored Data. The 'baking powder' can be offer to three customers ID: two as a new-offer and one as a re-buy. In the transactional history of the ID customers 8248 and 8285, bounded by the Maximum Recency property, there is no record that either customer have ever previously bought 'baking powder'. Note that they have 50% of probability to buy the 'baking powder'. This probability is given as a confidence measure because, in all transactions analyzed by the Bundles node, customers with the same historic patterns also bought 'baking powder' in 50% of the transactions. The other customer with ID 8407 besides sharing a similar historic pattern, it has previously bought 'baking powder' and it is also a potential buyer of this item, hence the offer type is now re-buy.



Scored Data

Item Name	Confidence (%)	ID	Offer Type
baking powder	50	8248	New Offer
baking powder	50	8285	New Offer
baking powder	50	8407	Re-buy
other vegetables	63.63636	1415	New Offer
other vegetables	63.63636	2750	Re-buy
other vegetables	63.63636	[illegible]	Re-buy

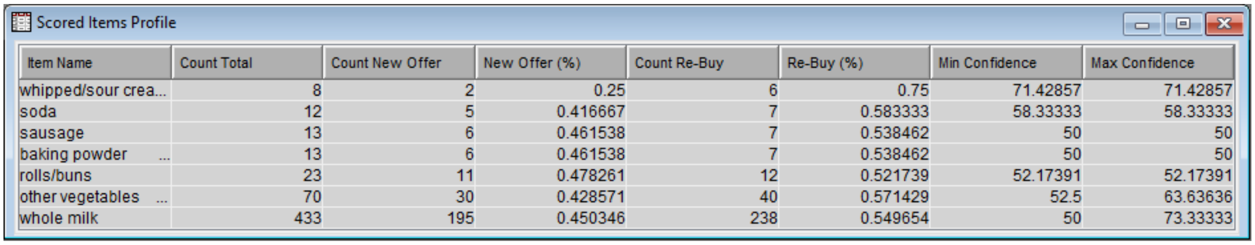
Display 14: Scored Data output example

Scored Items Profile

Profiles the potential of customers by type of offer for each item. Its columns are described as follows:

1. Item Name: Depict the name of item.
2. Taxonomy Level: Indicates the hierarchy of the item. This column appears when the taxonomy data is set and mapped on the Bundles Node.
3. Count Total: Number of potential customers for the item. It sums up the number of potential customers for new offer and potential customers for re-buy.
4. Count New Offer: Number of potential customers for new offer for the item.
5. New Offer (%): Calculate as the Count New Offer over Count Total.
6. Count Re-Buy: Number of potential customers for re-buy of the item.
7. Re-Buy (%): Calculate as the Count Re-Buy over Count Total.
8. Min Confidence: Indicates the minimum confidence found among all potential customers of the item.
9. Max Confidence: Indicates the maximum confidence found among all potential customers of the item.

The Display 15 depicts an output example of the Scored Data. The item 'whole milk' has 433 potential customers. Of these, 45% have no record that have ever previously bought 'whole milk, i.e., there are 195 customers eligible to buy this item as a new offer.



Scored Items Profile

Item Name	Count Total	Count New Offer	New Offer (%)	Count Re-Buy	Re-Buy (%)	Min Confidence	Max Confidence
whipped/sour crea...	8	2	0.25	6	0.75	71.42857	71.42857
soda	12	5	0.416667	7	0.583333	58.33333	58.33333
sausage	13	6	0.461538	7	0.538462	50	50
baking powder ...	13	6	0.461538	7	0.538462	50	50
rolls/buns	23	11	0.478261	12	0.521739	52.17391	52.17391
other vegetables ...	70	30	0.428571	40	0.571429	52.5	63.63636
whole milk	433	195	0.450346	238	0.549654	50	73.33333

Display 15: Scored Items Profile output example

Using the Next Best Offer Node

The Next Best Offer Node requires an input data source with a role of score that contains one ID variable and one target variable. The ID variable is used to group the target into transactions (baskets). The Target variable is a nominal variable that contains the item information. Optionally an extra numeric variable can be considered, usually the date, allowing the identification of the time sequence of transactions. This variable must have a role of sequence. This allows to do the Next Best Offer (NBO) based on Maximum Recency property of the Next Best Offer Node.

The maximum recency property specifies the maximum number of transactions to be selected for the Next Best Offer (NBO). There is an option to select ALL which considers all the historic transactional data of each ID. This option is the default even if there is no sequence variable. It ranges from 1 to 15. For instance, basing offers in the previous 3 transactions of the customer allows better alignment of offers and customer most recent behavior.

For further details of how to use this node, see the section Next Best Offer Node Properties.

Next Best Offer Node Properties

The following general properties are associated with the Next Best Offer Node:

- Node ID — The Node ID property displays the ID that SAS® Enterprise Miner™ assigns to a node in a process flow diagram. Node IDs are important when a process flow diagram contains two or more nodes of the same type. The first Next Best Offer Node added to a diagram will have a Node ID of NBO. The second Next Best Offer Node added to a diagram will have a Node ID of NBO2, and so on.

- Imported Data — The Imported Data property provides access to the Imported Data — Next Best Offer window. The Imported Data — Next Best Offer window contains a list of the ports that provide data sources to the Next Best Offer Node. Select the … button to the right of the Imported Data property to open a table of the imported data.

  If data exists for an imported data source, you can select the row in the imported data table and click one of the following buttons:

  - Browse to open a window where you can browse the data set.

  - Explore to open the Explore window, where you can sample and plot the data.

  - Properties to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contain summary information (metadata) about the table and variables.

- Exported Data — The Exported Data property provides access to the Exported Data - Next Best Offer window. The Exported Data - Next Best Offer window contains a list of the output data ports that the Next Best Offer Node creates data for when it runs. Select the … button to the right of the Exported Data property to open a table that lists the exported data sets.

  If data exists for an imported data source, you can select the row in the imported data table and click one of the following buttons:

  - Browse to open a window where you can browse the data set.

  - Explore to open the Explore window, where you can sample and plot the data.
  - Properties to open the Properties window for the data source. The Properties window contains a Table tab and a Variables tab. The tabs contain summary information (metadata) about the table and variables.
- Notes — Select the … button to the right of the Notes property to open a window that you can use to store notes of interest, such as data or configuration information.

Train Properties

- Maximum Recency — Specifies the maximum number of transactions to be selected for the Next Best Offer (NBO). This option only has effect when a sequence variable is found in the input data source with a role of score. The minimum value is 1, which means the NBO is based on the most recent transaction of each ID. If this value is 3 the NBO considers at most 3 recent transactions of each ID. This value ranges from 1 to 15. There is an option to select ALL which considers all the historic transactional data of each ID. This option is the default even if there is no sequence variable.

Status Properties

The following status properties are associated with this node:

- Create Time — displays the time at which the node was created.
- Run ID — displays the identifier of the node run. A new identifier is created every time the node runs.
- Last Error — displays the error message from the last run.
- Last Status — displays the last reported status of the node.
- Last Run Time — displays the time at which the node was last run.
- Run Duration — displays the length of time of the last node run.
- Grid Host — displays the grid server that was used during the node run.
- User-Added Node — specifies if the node was created by a user as a SAS® Enterprise Miner™ extension node.

CONCLUSION

This paper describes how to expand the usage of the SAS® Enterprise Miner™ by implementing product bundling and next best offer in a seamless way using two extension nodes. The Bundle Node is aimed to retail, but it can be adapted to a wide range of industries and market channels. Both nodes are particularly appropriated to industries where the customer ID is recorded in the transaction data, this enables personalization of the marketing campaign, by offering the right product to the right customer.

REFERENCES

SAS Institute Inc. 2017. SAS® Enterprise Miner™ 14.3 Extension Nodes: Developer's Guide Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2017. SAS® Enterprise Miner™ 14.3: Reference Help. Cary, NC: SAS Institute Inc.

Sarma, Kattamuri S., PhD. 2013. Predictive Modeling with SAS ® Enterprise Miner ™: Practical Solutions for Business Applications, Second Edition. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2011. SAS ® Certification Prep Guide: Base Programming for SAS ® 9, Third Edition. Cary, NC: SAS Institute Inc.

SAS Institute Inc. 2014. SAS ® Certification Prep Guide: Advanced Programming for SAS ® 9, Fourth Edition. Cary, NC: SAS Institute Inc.

Michael Hahsler, Kurt Hornik, and Thomas Reutterer (2006) Implications of probabilistic data modeling for mining association rules. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nuernberger, and W. Gaul, editors, From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization, pages 598–605. Springer-Verlag.

ACKNOWLEDGMENTS

Special thanks to the following individuals who provided their support to ensure that this paper was written: Jacqueline Fraga, Érica Albuquerque, Luisa Rosa and Vasco Jesus.

CONTACT INFORMATION

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

Bruno Nogueira Carlos
brunoncarlos@hotmail.com
bruno.nogueira@youmanmod.com
www.youmanmod.com