Market Basket Analysis

- Purpose
- Benefits
- Method
- Results
- Limitations
- Performing Market Basket Analysis
    - Virtual Items
    - Support
    - Confidence
    - Improvement
    - Taxonomies
    - Using Multiple Items
- Using the Results
- Store Layout Changes
- Product Bundling

Purpose

Market Basket Analysis is one of the most common and useful types of data analysis for marketing. The purpose of market basket analysis is to determine what products customers purchase together; it takes its name from the idea of customers throwing all their purchases into a shopping cart (a "market basket") during grocery shopping. Knowing what products people purchase as a group can be very helpful to a retailer or to any other company. A store could use this information to place products frequently sold together into the same area, while a catalog or World Wide Web merchant could use it to determine the layout of their catalog and order form. Direct marketers could use the basket analysis results to determine what new products to offer their prior customers.

In some cases, the fact that items sell together is obvious – every fast-food restaurant asks their customers "Would you like fries with that?" whenever they go through the drive-through window. However, sometimes the fact that certain items would sell well together is far from obvious. A well known example is that a supermarket performing a basket analysis discovered that diapers and beer sell well together on Thursdays. Though the result does make sense – couples stocking up on supplies for themselves and for their children before the weekend starts – it's not the sort of thing that someone would normally think of. The strength of market basket analysis is that by using computer data mining tools,

it's not necessary for a person to think of what products consumers would logically buy together – instead, the customers' sales data is allowed to speak for itself. This is a good example of data-driven marketing.

Once it is known that customers who buy one product are likely to buy another, it is possible for the company to market the products together, or to make the purchasers of one product the target prospects for another. If customers who purchase diapers are already likely to purchase beer, they'll be even more likely to if there happens to be a beer display just outside the diaper aisle. Likewise, if it's known that customers who buy a sweater from a certain mail-order catalog have a propensity toward buying a jacket from the same catalog, sales of jackets can be increased by having the telephone representatives describe and offer the jacket to anyone who calls in to order the sweater. By targeting customers who are already known to be likely buyers, the effectiveness of marketing is significantly increased – regardless of if the marketing takes the form of in-store displays, catalog layout design, direct offers to customers. This is the purpose of market basket analysis – to improve the effectiveness of marketing and sales tactics using customer data already available to the company.

Benefits

Knowing which products sell together can be very useful to any business. The most obvious effect is the increase in sales that a retail store can achieve by reorganizing its products so that things that sell together are found together. This facilitates impulse buying and helps ensure that customers who would buy a product don't forget to buy it on account of not having seen it. In addition, this has the side effect of improving customer satisfaction – once they've found one of the items they want, the customer doesn't have to look all over the store for something they want to buy. Their other purchases are already located close-by. World Wide Web or catalog merchants get the same benefit, by conveniently organizing their catalog or Web site so that items that sell together are found together.

Outside of the store environment, basket analysis provides different benefits, though equally useful ones. For a direct marketer, it is far preferable to market to existing customers, which are known to buy products and have a history with the company. The company already has these people in its database, and knows a significant amount of information about them. After running a basket analysis, a direct marketer can contact its prior customers with information about new products that have been shown to sell well with the products they've already bought; chances are, they'll be interested. In addition, even when making sales to new customers, telephone representatives can offer buyers of a product discounts on any other products they know sell with it, in order to increase the

size of the sale.

Finally, basket analysis has uses even outside the realm of marketing. It can be useful for operations purposes to know which products sell together in order to stock inventory. Running out of one item can affect sales of associated items; perhaps the reorder point of a product should be based on the inventory levels of several products, rather than just one. In addition, basket analysis can be used in any case where several different conditions lead to a result; by studying the occurrence of side effects in patients with multiple prescriptions, a hospital could find previously unknown drug interactions about which to warn patients.

There are several advantages to market basket analysis over other types of data mining – first of all, it is undirected. It is not necessary to choose a product that you want to focus on in order to run a basket analysis. Instead, all products are considered, and the data mining software reveals which products are most important to the analysis. In addition, the results of basket analysis are clear, understandable association rules that lend themselves to being immediately acted upon, and the individual calculations involved are simple.

Method

In order to perform market basket analysis, it is necessary to first have a list of transactions and what was purchased in each one. For simplicity, we will look at the example of convenience store customers, each of whom bought only a few items:

Transaction 1: Frozen pizza, cola, milk

Transaction 2: Milk, potato chips

Transaction 3: Cola, frozen pizza

Transaction 4: Milk, pretzels

Transaction 5: Cola, pretzels

Each customer purchased a different basket of items, and at first glance, there is no obvious relationship between any of the items purchased and any other item. The first step of a basket analysis, however, is to cross-tabulate the data into a table, allowing you to see how often products occurred together. For these five convenience store purchases, the table looks like this:

	Frozen Pizza	Milk	Cola	Potato Chips	Pretzels
Frozen Pizza	2	1	2	0	0
Milk	1	3	1	1	1
Cola	2	1	3	0	1
Potato Chips	0	1	0	1	0
Pretzels	0	1	1	0	2

The central diagonal of the chart shows how often each item was purchased with itself. Though this is significant for figuring some reliability statistics, it does not show how items sell together, and can be ignored for now. Look at the first row – out of the people who bought frozen pizza, one bought milk, two bought cola, and none bought potato chips or pretzels. This hints at the fact that frozen pizza and cola may sell well together, and should be placed side-by-side in the convenience store. Looking over the rest of the table, there is nowhere else that an item sold together with another item that frequently – this is probably an actual cross-selling opportunity. Compare this to the second row – of people who bought milk, one bought frozen pizza, one bought cola, one bought potato chips, and one bought pretzels. It seems milk sells well with everything in the store – there is probably not a good cross-selling opportunity with milk. This makes sense for a convenience store – people often come to a convenience store for the express purpose of buying milk, and will buy it regardless of anything else they're looking for.

Results

In the real world, there would usually be more than five products, and you would always have more than five transactions to look at. As a result, the distinction between products that sell well together and products that do not would be much sharper. Also, a market basket analysis of large amounts of data would be performed using data mining software, rather than being entered into a table by hand as we have done. Megaputer Intelligence offers a market basket analysis algorithm as part of its PolyAnalyst 4.0 integrated data mining suite.

The results of market basket analysis are particularly useful because they take the form of immediately actionable association rules . These are rules in the form of "if condition then result." For instance, from the above table, we could derive the association rules:

If a customer purchases Frozen Pizza, then they will probably purchase

Cola.

If a customer purchases Cola, then they will probably purchase Frozen Pizza.

These rules allow a store to immediately know that promotions involving frozen pizza and cola will pay off. Whether it's placing the cola display right next to the frozen pizza, advertising the two products together, or putting cola discount coupons on frozen pizza boxes, the convenience store will probably be able to increase sales of both items through directed marketing. And unlike most promotions, this promotion is almost sure to pay off – the convenience store has the data to back it up before even beginning the campaign.

This is an example of the best kind of market basket analysis result – the actionable. Unfortunately, there are two other kinds of association rules sometimes generated by market basket analysis – the trivial and the inexplicable . A trivial rule is one that would be patently obvious to anyone with some familiarity with the industry at hand. For instance, the discovery that hot dog buns sell well with hot dogs would not surprise the owner of a grocery store, and would in fact not be at all useful for promotion purposes – people will buy hot dog buns with their hot dogs regardless of any marketing campaign encouraging them to do so. Another example of a trivial rule would be the discovery that people who purchase an extended warranty for a television generally purchase a television – there would be no way for them to buy the warranty otherwise. The data mining software can only point out which items sell well together; it is up to the user to use his or her own business knowledge determine which rules are valuable to the business.

Finally, market basket analysis occasionally produces inexplicable rules. These rules are not obvious, but also don't lend themselves to immediate marketing use. An example of this type of rule is one hardware store chain's discovery that toilet rings sell very well only when a new hardware store is opened. There is no obvious reason for this – why do people only need toilet rings when a new store opens? In addition, while the company could offer a sale on toilet rings during new store openings, it's hard to tell whether or not this will be a successful promotion, since it's rather mysterious why they sell better at new openings at all. An inexplicable rule is not necessarily useless, but its business value is not obvious and it does not lend itself to immediate use for cross-selling.

Limitations

Though an useful and productive type of marketing data mining, market basket analysis does have a few limitations. The first is the kind of data needed to do an effective basket analysis – it is necessary to have a large number of real transactions to get meaningful data, but the data's accuracy is compromised if all of the products do not occur with similar frequency. Thus, in our convenience store example, if milk is sold in almost every transaction, but glue only sells once or twice per month, putting both of them into the same basket analysis will probably generate results that look impressive without being statistically significant – acting on these results might not actually benefit profitability. With only one or two glue customers, the data mining software will be able to very confidently state what sells well with glue – but this may only be true for the one or two customers analyzed. However, this limitation can be overcome by classifying items into a taxonomy as described in the next section.

Second, market basket analysis can sometimes present results that are actually due to the success of previous marketing campaigns. If the convenience store had always been putting cola discount coupons on the frozen pizza, the fact that cola and frozen pizza sell well together may come as no surprise to them – it does not give any new information, just show that previously existing marketing campaigns are already working. In fact, the previous campaign may even be overshadowing a real relationship – perhaps people would normally prefer to buy beer with pizza, but are only buying the cola because of the discount. In this case, the convenience store is missing out on what could be a better promotion.

Performing Market Basket Analysis

Virtual Items

Sometimes, a marketer wants to consider more than just which items sell together in developing their promotions. It may be important to know which items sell better to families with children, or to repeat customers, or to new customers. In this case, the sales data can be augmented with the addition of virtual items. A virtual item is not a real item being sold, but is treated as one by the data mining software. So if a new customer calls a catalog and orders a sweater and a jacket, this can be entered into the database as:

Item 1: Sweater

Item 2: Jacket

Item 3: (new customer)

Thus, when the data mining software is used to determine which items sell well

together, it may discover that some items sell particularly well with the (new customer) virtual item. This could tell the catalog company which items are so interesting as to lure new customers to their company, as opposed to only selling to long-time catalog buyers. By using virtual items, data about the customers themselves, or which store the items sold at, or which sales representative sold the item can be considered in the analysis without changing the method by which it is performed. By adding the store number or sales representative number as a virtual item, patterns can be found that exist only in certain places or are brought out by certain salespeople. As far as the data mining software is concerned, (new customer), or any virtual item, is a real item like any other.

Virtual items are also useful for testing the effects of promotions. By adding virtual items to represent promotions or discounts, it is possible to see how these affect cross-selling. These can also be used to compare urban and suburban sales, seasonal or time-of-day differences, or gift-wrapped purchases versus those that people bought for themselves.

Support

Market basket analysis can output any number of association rules, but only the best rules should be used for developing marketing campaigns. There are three measures of the quality of an association rule: support, confidence, and improvement.

Support is the percentage of records containing the item combination compared to the total number of records. So in our example series of transactions:

Transaction 1: Frozen pizza, cola, milk

Transaction 2: Milk, potato chips

Transaction 3: Cola, frozen pizza

Transaction 4: Milk, pretzels

Transaction 5: Cola, pretzels

	Frozen Pizza	Milk	Cola	Potato Chips	Pretzels
Frozen Pizza	2	1	2	0	0

Milk	1	3	1	1	1
Cola	2	1	3	0	1
Potato Chips	0	1	0	1	0
Pretzels	0	1	1	0	2

The support for the rule "If a customer purchases Cola, then they will purchase Frozen Pizza" is 40%. There are 5 total records, and 2 of them include both Cola and Frozen Pizza. Note that support considers only the combination, and not the direction – the support for the rule "If a customer purchases Frozen Pizza, then they will purchase Cola" is also 40%. Support can also be used to measure a single item – for instance, the support for the item "Milk" is 60%, since it occurs in 3 of the 5 records. Measuring the support of a single item is where the central diagonal of the table can be useful.

Confidence

Support, however, is an incomplete measure of the quality of an association rule. Is the 20% support for support for the combination of potato chips and milk a good rule? This could mean that 20% of all customers buy both potato chips and milk, and no one buys milk without also buying potato chips. In that case, it would be a good rule. But what if 100% of customers buy milk and only 20% of those buy potato chips? In this case, it's not a good rule, even though support is still 20%. The fact that a customer bought milk doesn't really tell you whether or not they'll buy potato chips – everyone bought milk. What is needed is a measure of how confident we can be, given that a customer has purchased one product, that they will also purchase another product.

Confidence provides this measure. The confidence of an association rule is the support for the combination divided by the support for the condition. For example, the rule "If a customer purchases Milk, then they will purchase Potato Chips" has a confidence of 33%. The support for the combination (Potato Chips + Milk) is 20%, occurring in 1 of the 5 transactions. However, the support for the condition (Milk) is 60%, occurring in 3 of the 5 transactions. This gives a confidence of (20% / 60%) = 33%.

Note also that confidence is directional – the confidence of the rule "If a customer purchases Potato Chips, then they will purchase Milk"" is (20% / 20%) = 100%. However, this rule is based on only one transaction! Thus, like a high support, a high confidence alone does not indicate that a rule is necessarily a good one. This also shows what happens when certain items with extremely low sales are thrown into a basket analysis with high sales – one customer's purchase

of two items together can create an extremely high-confidence rule that may not mean much. This problem is overcome by using taxonomies, as described later in this paper.

Improvement

Both support and confidence must be used to determine if a rule is valid. However, there are times when both of these measures may be high, and yet still produce a rule that is not useful. Take these results for example:

Convenience store customers who buy orange juice also buy milk with a 75% confidence. The combination of milk and orange juice has a support of 30%.

This at first sounds like an excellent rule, and in most cases, it would be. It has very high confidence and very high support. However, what if convenience store customers in general buy milk 90% of the time? In that case, orange juice customers are actually less likely to buy milk than customers in general. Thus, a third measure of the accuracy of market basket analysis is needed – improvement. This is defined as:

$$\frac{\text{Support (Condition + Result)}}{\text{Support(Condition)} * \text{Support(Result)}}$$

This can also be defined as the confidence of the combination of items divided by the support of the result. So in our milk example, assuming that 40% of the customers buy orange juice, the improvement would be:

$$\frac{30\%}{40\% * 90\%}$$

Which is 0.83 – an improvement of less than 1. Any rule with an improvement of less than 1 does not indicate a real cross-selling opportunity, no matter how high its support and confidence, because it actually offers less ability to predict a purchase than does random chance.

As a side note, it is possible to negate a rule that has an improvement of less than 1 and thereby produce a rule with an improvement greater than 1. For example, we could change the above rule into "Customers who buy orange juice do not buy milk with a 25% confidence. The purchase of orange juice without milk has a support of 10%." This rule offers an improvement of 1.21. Unfortunately, as is probably obvious from that rule, negated rules are usually not very useful for marketing. Knowing that customers who buy one product will not buy another one is usually not helpful for marketing products to

customers or placing the products in a store or catalog.

Taxonomies

The most common obstacle to performing a good market basket analysis is the presence of low-support items. The fact that 100% of purchasers of mushroom pizza also buy broccoli seems like a good rule, unless only once person has ever purchased a mushroom pizza. However, data mining software will produce such results, since they have a very high confidence and improvement, probably much higher than those of any item that sold a larger number of times.

There are two ways to deal with this problem. One way is to create a support threshold. Any combination that has support below a certain percentage will be dropped from the analysis. If the support threshold was set at 5%, it follows that since the support of a combination is always less than the support of any single item in it, all items with a support of less than 5% can also be dropped from the analysis. Thus, in this case, mushroom pizza would not even be considered in the analysis, and would thus not produce its 100%-confidence, high-improvement rule.

Unfortunately, the support threshold method has a major disadvantage – it is eliminating some potentially valuable data from consideration. This brings us to the best way to deal with low-support items: the creation of a taxonomy. A taxonomy is an orderly hierarchy of items and item categories, dividing things such that each item put into the basket analysis occurs with a similar level of support. This is done by aggregating low-support items into groups and analyzing the group as a single item, while breaking down high-support items into smaller units for analysis.

For example, look at the following list of items and their support levels:

Grape Popsicles 1%

Cherry Popsicles 3%

Orange Popsicles 3%

Lime Popsicles 2%

Frozen pizzas 70%

Obviously, this will not do for basket analysis. If someone buys a grape Popsicle and any other item, it will show an extremely high confidence and improvement

rule linking grape Popsicles with whatever else the person bought. To come up with a confidence or improvement this high, frozen pizza buyers would have to buy another item almost every time.

The solution to this is, obviously, to aggregate some items and break down others. The items that should be used in the analysis would go more like:

> Popsicles (all) 9%
>
> Frozen pizzas (pepperoni) 12%
>
> Frozen pizzas (cheese) 15%
>
> Frozen pizzas (Supreme) 11%
>
> Frozen pizzas (sausage) 10%
>
> Frozen pizzas (combination) 15%
>
> Frozen pizzas (vegetarian) 7%

By combining all the popsicles into a single item, while breaking the frozen pizzas down into smaller items, comparable support levels were found for all of the items. The support levels do not have to be the same, but if one item has support more than an order of magnitude above another item, the smaller-support items will probably dominate the association rules produced.

Taxonomies can be created aggregating items all the way up to "Frozen Food" or splitting them down to the UPC or SKU level. By finding comparable support levels for all items, you ensure the production of association rules whose confidence and improvement can be meaningfully used for comparison.

Using Multiple Items

Thus far, we have only looked at association rules involving two items. However, sometimes better rules emerge when more than two items are considered:

> If a customer purchases a plant and a clay pot, then they will purchase soil.
>
> If a customer purchases glue and scissors, then they will purchase paper.

Though these are trivial rules, they also illustrate that while sometimes a single item will not give much insight into a customer's future products, a pair of items might. The same goes for higher numbers of items.

Performing a basket analysis that considers higher numbers of items in groups is done iteratively – first pairs are found, then sets of three, then four, and so on. The number of calculations required to perform the analysis varies exponentially with the number of products to be considered simultaneously – the number of calculations for n items is proportional to the number of items to be considered at a time raised to the n power. As a result, a pruning method has been developed to minimize calculation time by eliminating items as the number of items to be considered simultaneously increases. To perform a multidimensional basket analysis, a minimum support threshold, say 2%, must be set. The data mining software first eliminates all items that have less support than this minimum threshold, then does an analysis comparing only pairs of items and generates a set of association rules.

At this point, the second round of pruning occurs. Any combination of two items that, as a pair, have a support less than the minimum threshold are eliminated from consideration as conditions of an association rule. Then, these pairs of items are checked against all the items in the analysis (as results) and another set of association rules are generated.

This process continues, next eliminating all sets of three items that as a group fall below the minimum support threshold. In some environments, such as a convenience store, it is quite possible that customers buy so few items at a time that no rules involving more than two or three items will ever have the minimum support necessary to be considered significant; in an environment like a grocery store, where customers buy over a hundred items at a time, rules of 10-12 items may be significant.

Performing basket analyses considering more than two items at a time results in the development of multi-dimensional tables that can be difficult to visualize. However, the use of data mining software allows meaningful rules to be found in this data despite the difficulties in representing it.

Using the Results

Store Layout Changes

The results of market basket analysis can be used by stores to change their layout in ways that improve profitability. If the basket analysis shows that light bulbs and gardening tools sell well together in a hardware store, the obvious

response is to put the light bulbs next to the gardening aisle. However, a better response might be to put a shelf of the store brand of light bulbs, a very high-profit item for the store, next to the gardening aisle, leaving the rest of the light bulbs where they are. By making it most convenient for the customer to buy high-profit items for the store, the store owner can maximize profit. The market basket analysis shows that this tactic will probably work, since customers will already be looking to buy the item.

This same tactic is equally valid for "stores" that take some form other than the supermarket floor – anywhere that a customer browses for items is appropriate for reorganization based on market basket analysis. A catalog or web page can also be reorganized so that customers who are likely to buy a certain product have their attention directed to high-profit items.

Product Bundling

For companies that don't have a physical storefront, like mail-order companies, Internet businesses, and catalog merchants, market basket analysis can be more useful for developing promotions than for reorganizing product placement. By offering promotions such that the buyers of one item get discounts on another they have been found likely to buy, sales of both items may be increased. In addition, basket analysis can be useful for direct marketers for reducing the number of mailings or calls that need to be made. By calling only customers who have shown themselves likely to want a product, the cost of marketing can be reduced while the response rate is increased.