Market Basket Analysis of Sales Data

“The performed Market Basket Analysis was very useful to the client. I would like to continue working together with Megaputer on other CTP customers' projects.”

Olof Goransson, Senior Data Consultant, CTP

Case Prepared By: Grant Bugher, Senior Data Analyst, Megaputer intelligence, Inc.

The goal of the CTP analysis was to determine how likely a customer would purchase a specific product, given the knowledge of what other items are already in this customer's "shopping basket". We need to find the percentage chance of purchasing this specific additional product. In addition, measures of statistical significance needed to be calculated to make sure that these percentages were greater than random chance. Finally, the most significant rules needed to be extracted. Meaningful conclusions were found in the data, and these are provided after a description of the analysis.

The data provided was in the form of sales data on 1,175 customers, listing the sales volume for each of 255 products for each customer. PolyAnalyst 4.0 includes as an add-in module the cross-sell analysis algorithm that could handle this data more easily, but for now, we ran the analysis through more conventional means. The analysis that follows is only the simplest analysis of what we can do with the data; it is also possible to find cross-sell opportunities involving 3, 4, or more products sold together, as well as to do some analysis involving the sales volumes.

First, the data was converted and imported into an Excel spreadsheet. The sales volumes, not being useful for the type of analysis being currently used, were converted to Boolean data -- only what products each customer purchased, and not how much of them, was considered. First, we tabulated the frequencies with which each item was sold with each other item -- the result was a 255 by 255 spreadsheet of frequencies.

Cross-sell analysis works best when every item occurs approximately the same number of times in the data. With this data, this was not the case, with some items occurring more than 400 times and other items occurring as few as 2. The best way to deal with this frequency variation would be by creating a taxonomy -- grouping less-frequent items together into larger categories, while splitting more-frequent items up into smaller subgroups. The analysis would then be run on similarly sized groups. However, since we did not have any way to intelligently categorize the products (all we have is an unintelligible product code), all items that occurred in less than 5% of the 1,175 transactions (59 transactions) were simply eliminated from further analysis. This reduced the pool of products to 95.

From the frequency data, the probability that a given transaction will include each product was calculated. Coupled with the data listing how often each product sold with each other product, this allowed us to generate a confidence score for each combination of two products. The confidence score is an indicator that if a customer bought the first product in the pair, of how likely they are to buy the second product in

the pair. This is the first major output of this cross-sell analysis -- items with a high confidence are good opportunities for cross-selling.

Using the confidence and frequency measures, another output can be generated -- the improvement score. The improvement score shows how much better the prediction is than random chance. It indicates how much more likely a customer is to buy the second product given that he bought the first product than he would be to buy the second product given no other information about him. This is the second major output of the cross-sell analysis.

If an improvement score is less than 1, it indicates that regardless of the confidence measure, there is not really a cross-sell opportunity, because the customer would be more likely to buy the other product by simple random chance. Likewise, if a confidence score is less than 50%, there is no cross-sell opportunity, since purchasing the first product indicates less than a 50% chance that the second product will be purchased. In addition, the higher the confidence score is, the more likely that cross-selling will occur.

To get actionable rules out of this data, all cells in the spreadsheet exhibiting a confidence less than 80% or an improvement ratio of less than 4.0 were eliminated. This left 31 cross-sell rules (combinations of two products in order) shown to be valid.

Combination	Improvement	Confidence	Support
60B -> 60D	12.095	88.52%	4.60%
66K -> 66M	6.528	81.11%	6.21%
66F -> 66M	6.770	84.13%	4.51%
14E -> 14I	5.458	85.94%	4.63%
85B -> 14I	5.197	81.82%	4.60%
02P -> 14I	5.119	80.60%	4.60%

The Support column indicates in what percentage of the transactions the combination occurs. Note that these rules are one-way -- though a purchase of Item 31F indicates an 87.10% chance that the customer will also purchase an 02C, the purchase of an 02C does not indicate a similar probability that the customer will purchase a 31F.

The following conclusions can be derived from this data:

The first is the rules themselves -- whenever one of the products on the left side of the rule is sold, a customer should be offered the product on the right side of the rule, as they are very likely to purchase it. The confidence levels in these rules are more than 80%, so they can be acted on reliably.

Some products (14I, 14A, 11F) show up with very high confidence with many other purchases. These are probably some sort of support item that should be bundled with the products they are frequently sold with.

In addition, the sales of Item 60B not only indicated the sale of item 60D with very high confidence, but also showed an extremely high improvement (over 12) indicating that Item 60D is far more likely to sell in conjunction with 60B than on its own.

Items 66M, 79A, and 02C were also shown to reliably sell with two other items each. Bundling these items (66M with 66K and 66F, 79A with 75E and 73D, and 02C with 08B

and 31F) or encouraging their sale when their linked items are sold should improve their sales rate.

The highest confidence rules took the form of a low-volume product leading to the purchase of a high-volume product. For instance, in the rule 14E -> 14A, 14E occurs in only 64 transactions (near our minimum allowed of 59), while 14A occurs in 198 transactions. This pattern is evident in most of the rules discovered.

In addition, the following improvements in the data would enable better conclusions:

A taxonomy should be developed, allowing lower-volume items to be grouped together and higher-volume items to be split so that each item or category occurs approximately the same number of times in the data. This will improve the quality of the cross-sell analysis.

A unit price for each item (or average unit price for categories) should be devised so that the number of units sold can be determined. At present, it is impossible for us to know if a high volume indicates the sales of many inexpensive items or a single very expensive item.

Finally, keep in mind that market basket analysis sometimes produces trivial or non-actionable rules. For instance, an analysis of the sales of a computer store may reveal that extended warranties are generally purchased when computers are purchased -- though true, it is also obvious, since no one would have any reason to buy an extended warranty if they had not also bought a computer. Unfortunately, since we had only product code numbers and no indication of what the products actually were, we have no way of identifying if any of these rules are trivial or if they are all valid, useful rules. They should be evaluated by someone familiar with the products themselves.