Introduction

Association rules are popular in marketing for cross-selling products associated with an item that a consumer is considering. In association rules, the goal is to identify item clusters in transaction type databases. Association rule discovery in marketing is termed "market basket analysis" and is aimed at discovering which groups of products tend to be purchased together. These items can then be displayed together, offered in post-transaction coupons, or recommended in online shopping. Put simply, association rules, or *affinity analysis*, constitute a study of "what goes with what." This method is also called *market basket analysis* because it originated with the study of customer transactions databases to determine dependencies between purchases of different items. Association rules are heavily used in retail for learning about items that are purchased together, but they are also useful in other fields. For example, a medical researcher might want to learn what symptoms appear together. In law, word combinations that appear too often might indicate plagiarism.

Discovering Association Rules in Transaction Databases

The availability of detailed information on customer transactions has led to the development of techniques that automatically look for associations between items that are stored in the database. An example is data collected using bar-code scanners in supermarkets. Such *market basket databases* consist of a large number of transaction records. Each record lists all items bought by a customer on a single-purchase transaction. Managers are interested to know if certain groups of items are consistently purchased together. They could use such information for making decisions on store layouts and item placement, for cross-selling, for promotions, for catalog design, and for identifying customer segments based on buying patterns. Association rules provide information of this type in the form of "if– then" statements. These rules are computed from the data; unlike the if–then rules of logic, association rules are probabilistic in nature.

Definition of terms

We use the term *antecedent* to describe the IF part, and *consequent* to describe the THEN part. In association analysis, the antecedent and consequent are sets of items (called *item sets*) that are disjoint (do not have any items in common). Note that item sets are not records of what people buy; they are simply possible combinations of items, including single items.

The support of a rule is simply the number of transactions that include both the antecedent and consequent item sets. It is called a support because it measures the degree to which the data "support" the validity of the rule. The support is sometimes expressed as a percentage of the total number of records in the database. What constitutes a frequent item set is therefore defined as an item set that has a support that exceeds a selected minimum support, determined by the user.

Support and Confidence

In addition to support, there is another measure that expresses the degree of uncertainty about the if–then rule. This is known as the *confidence of* the rule. This measure compares the co-occurrence of the antecedent and consequent item sets in the database to the occurrence of the antecedent item sets. Confidence is defined as the ratio of the number of transactions that Include all antecedent and consequent item sets (namely, the support) to the number of transactions that include all the antecedent item sets:

$$\text{Confidence} = \frac{\text{no. of transactions with both antecedent and consequent itemsets}}{\text{no. of transactions with antecedent itemset}}.$$

For example, suppose that a supermarket database has 100,000 point-of-sale transactions. Of these transactions, 2000 include both orange juice and (over-the-counter) flu medication, and 800 of these include soup purchases. The association rule "IF orange juice and flu medication are purchased THEN soup is purchased on the same trip" has a support of 800 transactions (alternatively, 0.8% = 800/100,000) and a confidence of 40% (=800/2000).

Lift Ratio

A better way to judge the strength of an association rule is to compare the confidence of the rule with a benchmark value, where we assume that the occurrence of the consequent item set in a transaction is independent of the occurrence of the antecedent for each rule. In other words, if the antecedent and consequent item sets are independent, what confidence values would we expect to see? Under independence, the support would be:

P(antecedent AND consequent) = P(antecedent) × P(consequent), and the benchmark confidence would be:

P(antecedent) × P(consequent)/ P(antecedent) = P(consequent)

The estimate of this benchmark from the data, called the *benchmark confidence value* for a rule, is computed by:

$$\text{Benchmark confidence} = \frac{\text{no. of transactions with consequent item set}}{\text{no. of transactions in database}}$$

We compare the confidence to the benchmark confidence by looking at their ratio: this is called the *lift ratio* of a rule. The lift ratio is the confidence of the rule divided by the confidence, assuming independence of consequent from antecedent:

Lift ratio = confidence/benchmark confidence.

A lift ratio greater than 1.0 suggests that there is some usefulness to the rule. In other words, the level of association between the antecedent and consequent item sets is higher than would be expected if they were independent. The larger the lift ratio, the greater the strength of the association. To illustrate the computation of support, confidence, and lift ratio for the cellular phone faceplate example, we introduce an alternative presentation of the data that is better suited to this purpose

Application of Association Rule

Catalog Cross-Selling
CatalogCrossSell.csv is the dataset for this case study.

Background
Exeter, Inc. is a catalog firm that sells products in a number of different catalogs that it owns. The catalogs number in the dozens, but fall into nine basic categories:
1. Clothing
2. Housewares
3. Health
4. Automotive
5. Personal electronics
6. Computers
7. Garden
8. Novelty gift
9. Jewelry

The costs of printing and distributing catalogs are high. By far the biggest cost of operation is the cost of promoting products to people who buy nothing. Having invested so much in the production of artwork and printing of catalogs, Exeter wants to take every opportunity to use them effectively. One such opportunity is in cross selling—once a customer has "taken the bait" and purchases one product, sell them another while their attention is at the peak.
Such cross-promotion might take the form of enclosing a catalog in the shipment of a purchased product, together with a discount coupon to induce a purchase from that catalog. Or, it might take the form of a similar coupon sent by e-mail, with a link to the web version of that catalog.

But which catalog should be enclosed in the box or included as a link in the e-mail with the discount coupon? Exeter would like it to be an informed choice—a catalog that has a higher probability of inducing a purchase than simply choosing a catalog at random.

Task
Using the dataset *CatalogCrossSell.csv*, an association rules analysis will be performed, and comments on the results will be given. There will also be a very rough estimate of the extent to which this will help Exeter make an informed choice about which catalog to cross-promote to a purchaser.

S/N	Customer Number	Clothing Division	Housewares Division	Health Products Division	Automotive Division	Personal Electronics Division	Computers Division	Garden Division	Novelty Gift Division	Jewelry Division
0	11569	0	1	1	1	1	0	0	1	0
1	13714	0	1	1	1	1	0	1	1	1
2	46391	0	1	1	1	1	0	1	1	1
3	67264	0	0	1	1	1	0	1	1	0
4	67363	0	0	1	0	1	0	1	1	0
5	72553	0	1	1	1	1	0	1	1	1
6	79814	0	1	1	0	1	0	1	0	0
7	80903	0	1	1	0	1	0	0	1	0
8	91439	0	0	1	1	1	0	1	0	1
9	96701	0	1	1	1	1	0	1	1	1
10	98517	1	1	1	1	1	0	1	1	1

The table above shows a subset of the Catalog Cross-sell data. It captures customer number and all the different catalog categories. For example, the catalogs bought by a customer with number 11569 are marked 1 and they are: housewares division, health products division, automotive division and novelty gift division catalogs. The ones marked 0 were not purchased.

Note: The data contains a total of 4998 transactions and each row represents a unique transaction of a customer.

The result after applying apriori algorithm to the dataset is shown below

S/N	Antecedents	Consequents	Support	Confidence	Lift
1	(Housewares Division)	(Health Products Division)	0.393557	1.000000	1.000000
2	(Personal Electronics Division)	(Health Products Division)	0.467387	1.000000	1.000000
3	(Garden Division)	(Health Products Division)	0.272109	1.000000	1.000000
4	(Novelty Gift Division)	(Health Products Division)	0.356943	1.000000	1.000000
5	(Jewelry Division)	(Health Products Division)	0.235494	1.000000	1.280252
6	(Personal Electronics Division, Housewares Division)	(Health Products Division)	0.235494	1.000000	1.280252

7	(Personal Electronics Division)	(Housewares Division)	0.235494	0.503853	1.280252
8	(Housewares Division)	(Personal Electronics Division)	0.235494	0.598373	1.280252
9	(Health Products Division, Housewares Division)	(Personal Electronics Division)	0.235494	0.598373	1.280252
10	(Health Products Division, Personal Electronics Division)	(Housewares Division)	0.235494	0.503853	1.280252
11	(Housewares Division)	(Health Products Division, Personal Electronics Division)	0.235494	0.598373	1.280252
12	(Personal Electronics Division)	(Health Products Division, Housewares Division)	0.235494	0.503853	1.280252

Explanation of result

Assume a minimum support of 20% of the total transaction as the benchmark and consider only antecedents and consequents with single items with lift ratio > 1. Apparently, the best catalog for Exeter to cross sell based on the stated conditions is Jewelry with Health Products.

If Jewelry Division (JD) is purchased, then with 100% confidence Health Products Division (HPD) will also be purchased. This rule has a lift ratio of 1.28 and a support of 23.5%.

If Personal Electronics Division (HD) is purchased, then with 50.4% confidence Housewares Division will also be purchased. This rule has a lift ratio of 1.28 and a support of 23.5%.

If Housewares Division (HD) is purchased, then with 59.8% confidence Housewares Division will also be purchased. This rule has a lift ratio of 1.28 and a support of 23.5%.

If JD division is sold for ₦50 and HPD is sold for ₦10, JD can be cross sold with HPD. If there is a discount of 25% on the sales of HPD for every customer that purchases JD, then it means there is a high probability Exeter will make ₦57.5 by cross-selling JD and HPD.

Generally, if we consider the antecedents and consequents above with lift ratio > 1, HPD has a higher probability of inducing a purchase when a discount is introduced.