Association Mining (Market Basket Analysis)

(adapted from Applied Analytics using SAS Enterprise Miner, SAS Institute, Cary, NC. 2010)

The **BANK** data set contains service information for nearly 8,000 customers. There are three variables in the data set, as shown in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Model Role** | **Measurement Level** | **Description** |
| **ACCOUNT** | ID | Nominal | Account Number  |
| **SERVICE** | Target | Nominal | Type of Service |
| **VISIT** | Sequence | Interval | Order of Product Purchase |

The **BANK** data set has over 32,000 rows. Each row of the data set represents a customer-service combination. Therefore, a single customer can have multiple rows in the data set, and each row represents one of the products he or she owns. The median number of products per customer is three.

The 13 products are represented in the data set using the following abbreviations:

ATM automated teller machine debit card

AUTO automobile installment loan

CCRD credit card

CD certificate of deposit

CKCRD check/debit card

CKING checking account

HMEQLC home equity line of credit

IRA individual retirement account

MMDA money market deposit account

MTG mortgage

PLOAN personal/consumer installment loan

SVG saving account

TRUST personal trust account

Your first task is to create a new project with an analysis diagram and data source for the **BANK** data set.

The **project** will contain all information about your analysis, an **analysis diagram** is where you construct your model, and a **data source** is how you associate your projects with SAS data sets.

**Step 1: Create a New Project**

1. Once you log into SAS Enterprise Miner, click on New Project.
2. Click “Next” for step 1 of the wizard. This indicates your project will be stored on the SAS server.
3. At Step 2, enter a project name. Call it “Associations.”



Then click “Next.”
4. Click “Next” again at Step 3 and click “Finish” at Step 4.

**Step 2: Create a New Diagram**

1. Right-click on “Diagrams” and select “Create Diagram”


2. Enter “BankAnalysis” for the diagram name. Then click “OK.”


3. You’ll see the blank workspace on the right side of your screen and a new diagram under the “Diagrams” folder:



**Step 3: Create the Data Source**

1. Right-click on “Data Sources” and select “Create Data Source.”


2. At Step 1, make sure “Metadata Repository” is selected and click “Next.”
3. At Step 2, click “Browse” and select the folder “Shared Data.” Then select the folder “Libraries.” Then select the folder “AAEM.” Finally, select the table called “BANK.” Click “OK.”

You should see the following screen:



Click “Next.”

1. Click “Next” through Steps 3, 4, and 5.
2. At Step 6, assign metadata to the table variables as shown below. Pay attention to the **Role** and the **Level** columns.



What you’re saying here is that each ACCOUNT is a customer case, and therefore the Role is ID – the account identifies the customer.

The SERVICE is what product each customer purchases during a visit. The SERVICE can be a checking account, a savings account, etc. The purchase of services is the subject of the associations, so its role is Target.

The VISIT indicates the order that the SERVICE was added to the ACCOUNT. This is important to determine if a customer was more likely to buy a certain service when they already have another service. So the role of VISIT is Sequence.

1. Click “Next” for Steps 7 and 8.
2. At Step 9, make sure the Name is “BANK” and the role is “Transaction.” Then click “Next.”



1. Click “Finish” at Step 10 to finish creating the data set.

**Step 4: Explore the Data Set**

1. Right-click on the BANK data source and select “Explore.” It will take a few seconds to bring up the next window.



1. Inspect the data window at the bottom of the screen (AAEM.BANK). This is the data set with 32,367 unique transactions from the bank – each one is a purchase of a service by a customer. However, this window only displays the first 10,000 rows. Scroll through the list and verify that there are 10,000 observations. Now take a look at the ones at the top of the list:



This says that customer 500026 first purchased a checking account, then a savings account, then two ATM debit cards.
2. Now inspect the Sample Statistics window:



From this you can associate the variables in the data set with their labels shown in the data window. You also see that the highest number of visits a customer had with the bank was 11, and that the bank’s average (mean) number of visits was 2.8. This means that the average customer has *about* three products with the bank.

1. Finally, let’s make a plot of the VISIT data. Under the Actions menu, select “Plot…”
2. Select Histogram and click “Next.”
3. Assign the VISIT variable to the Role of “X.” Then click “Finish.”



1. You’ll see the following plot:



This indicates that a lot of people had between 1 and 3 visits, and very few had more than 7.

1. Close the Explore/Plot window.

**Step 5: Perform the Association Analysis**

1. Drag the “BANK” data source into the diagram workspace.
2. Select the “Explore” tab and drag an “Association” tool () into the diagram workspace.
3. Connect the “BANK” node to the “Association” node by drawing a line between them with your mouse.



1. Select the “Association” node and examine its Properties panel on the left side of the screen.



1. Scroll down to find the “Export Rule by ID” property. Set it to “Yes” – this makes sure that the **Rule Description** table will be available for display in the Results window.


Here are some of the other options in the Properties panel:

* The **Minimum Confidence Level** specifies the minimum confidence level to generate a rule. The default level is 10%.
* The **Support Type** specifies whether the analysis should use the support count or support percentage property. The default setting is Percent.
* The **Support Count** specifies a minimum level of support to claim that items are associated (that is, they occur together in the database). The default count is 2.
* The **Support Percentage** specifies a minimum level of support to claim that items are associated (that is, they occur together in the database). The default frequency is 5%. The support percentage figure that you specify refers to the proportion of the largest single item frequency, and not the end support.
* The **Maximum Items** determine the maximum size of the item set to be considered. For example, the default of four items indicates that a maximum of four items will be included in a single association rule.

**🖉** If you are interested in associations that involve fairly rare products, you should consider reducing the support count or percentage when you run the Association node. If you obtain too many rules to be practically useful, you should consider raising the minimum support count or percentage as one possible solution.

We first will perform a basic market basket analysis without taking into account the sequence. That will tell us how likely someone who has product A also has product B. So for now, we do not need the sequence variable.

1. Right-click on the “Associations” node and select “Edit Variables.”



Make sure “VISIT” is set to “No” for the Use property. Then click “OK.”

1. Run the diagram from the Association node (right-click on the node and select “Run.” It will take about a minute to run.
2. View the results by clicking on the “Results…” button in the “Run Status” dialog:


3. Now you’ll see the “Results - Node: Association Diagram” window opens with the Statistics Plot, Statistics Line Plot, Rule Matrix, and Output windows visible (although they might be in a different order).



1. To view the actual association rules, from the View menu select Rules/Rule Description.



The rules are in descending order according to lift. The highest lift rule is checking and credit card implies check card. In other words, those customers that have a checking account and a credit card also have a check card.

This is not surprising given that these are complementary services – the check card service goes with a checking account and is backed by a credit card company. Notice the symmetry in rules 1 and 2. This is not accidental because lift is symmetric (in other words, if B is associated with A, then A is associated with B).

1. Now examine the rule matrix by expanding that window.



The rule matrix is a plot of every rule based on the combinations of bank products that appear on the left side (antecedents) and right side (consequents) of all rules. Each point in the matrix is a rule. The points are colored, based on confidence (blue = low confidence and red = high confidence). The rules with the highest confidence are in the column in the picture above.

Click on a rule (one of the dots) and then hover your mouse over it. You’ll see the left and right hand side of that rule. You can see that rules on the same horizontal line have the same left-hand side (antecedent), and the rules on the same vertical line have the same right-hand side (consequent). Note that the right hand side of the rule is “checking” for all rule combinations in the high confidence column (red).

1. Another good way to explore the rules found in the analysis is by viewing the Rules table. From the View menu select Rules/Rules Table.



You can see which rules have the highest lift, confidence, etc. Clicking on the column heading puts them in order by that value.

1. Another way to view these rules graphically is to use the plot wizard.
Click the Plot Wizard icon ().
2. Choose a scatter plot and then select the 3-D option:

and then select “Next.”
3. Select the roles “X,” “Y,” and “Z” for the variables SUPPORT, LIFT, and CONF. Like this:



1. Select **Finish** to create the plot.
2. Rearrange the windows to view the data and the plot simultaneously. Hold down the Control key and then click and drag your mouse on the graph to rotate it.



If you select a point on the graph is will highlight the corresponding rule in the table. If you select a rule in the table it will highlight the corresponding rule in the graph. You can also drag your mouse to draw a box around those high-lift rules on the right side of the graph. You’ll see the first six rules or so become highlighted in the rules table.

You can use this to examine high-lift rules and view their corresponding support and lift values.

1. Now close the Results window.

Sequence Analysis

In addition to the products owned by its customers, the bank is interested in examining the order in which the products are purchased. The sequence variable in the data set enables you to conduct a sequence analysis.

1. Add a second Association node to the diagram workspace and connect it to the “BANK” node.
2. Rename the new node SequenceAnalysis by right-clicking on the new node and selecting “Rename.” This time you will NOT exclude the Visit variable. You’re now including the sequence information – the order in which services are purchased – in the analysis.



1. Set “Export Rule by ID” to “Yes,” just like you did before.



1. Examine the Sequence section of the Properties panel.



The options in the Sequence panel enable you to specify the following properties (you won’t be changing anything for this analysis):

* **Chain Count** is the maximum number of items that can be included in a sequence. The default value is 3 and the maximum value is 10.
* **Consolidate Time** enables you to specify whether consecutive visits to a location or consecutive purchases over a given interval can be consolidated into a single visit for analysis purposes. For example, two products purchased less than a day apart might be considered to be a single transaction.
* **Maximum Transaction Duration** enables you to specify the maximum length of time for a series of transactions to be considered a sequence. For example, you might want to specify that the purchase of two products more than three months apart does not constitute a sequence.
* **Support Type** specifies whether the sequence analysis should use the Support Count or Support Percentage property. The default setting is Percent.
* **Support Count** specifies the minimum frequency required to include a sequence in the sequence analysis when the Sequence Support Type is set to Count. If a sequence has a count less than the specified value, that sequence is excluded from the output. The default setting is 2.
* **Support Percentage** specifies the minimum level of support to include the sequence in the analysis when the Support Type is set to Percent. If a sequence has a frequency that is less than the specified percentage of the total number of transactions, then that sequence is excluded from the output. The default percentage is 2%. Permissible values are real numbers between 0 and 100.
1. Run the diagram from the Sequence Analysis node and view the results.
2. From the View menu, select Rules/Rule description. You may have to expand the Rule column see the full rule. You do this by clicking on the right side of the column heading and dragging it a little to the right. The rules are in descending order of support (frequency of occurrence within the set).



Introducing the order of service acquisition into the analysis changes the type of information we gather from the analysis. We now see that it is most common for customers who have a checking account first to then open a savings account (RULE1). It is also very common for customers who already have a savings account are likely to get an ATM card (RULE3).

Performing a sequence analysis can give you different insight than a regular association rule analysis. We’re not just showing what things happen together, but also *why* they happen. However, order of occurrence has to be meaningful in the context of the problem.