IDC3931 Data Mining Lab:

Welcome to Data Mining! This handout is intended to provide you with information you will need to try your hands at data mining. We will build on your previous experience with your DataMart that you created on the Teradata server and you will utilize remote desktop from any windows machine in order to connect to your DataMart. The handout is divided into the following sections:

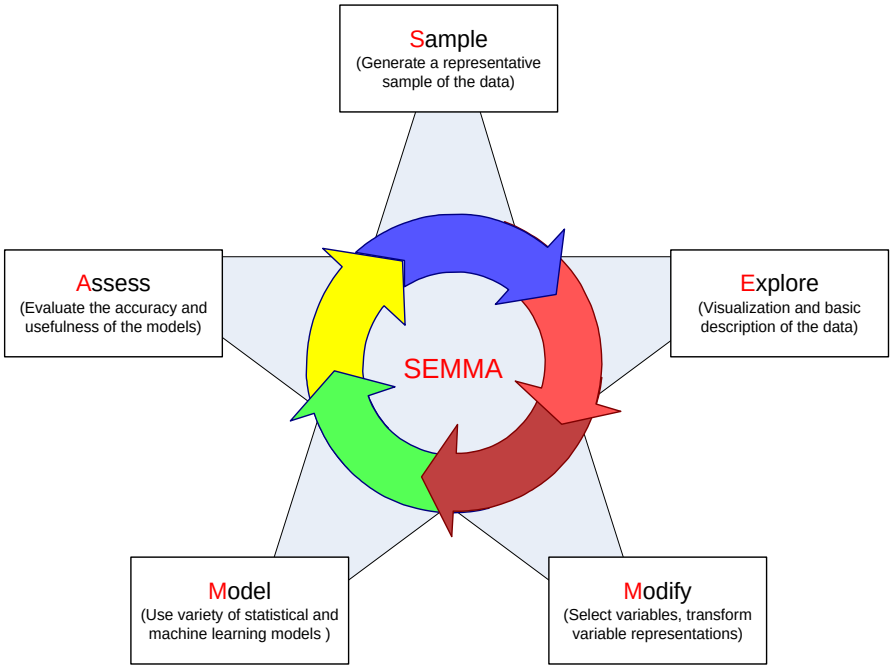
- **Basics of Data Mining** – this section provides you with the background information regarding the dataset you will be using
- **Introduction to Teradata Warehouse Miner(TWM)** – this section demonstrates how you can get set up and running with the Teradata Warehouse miner software. It includes basic topics creating a project, defining the appropriate data sources, and various Analysis models for TWM
- **Association Discovery** – this section covers details of how to prepare for, perform and interpret an association discovery analysis model using TWM.
- **Decision Tree** – this section covers details of how to prepare for, perform and interpret a Decision Tree analysis model using TWM.

Basics of Data Mining

You will use should use Teradata SQL Assistant and Teradata Warehouse Miner TWM for this homework. The main steps of data mining can be grouped under two main headings:

- **Data preparation –** entails the selection of the appropriate pieces of data from the raw dataset and formatting of them.
  - **Data Consolidation –** create your DataMart – did this for Assignment 1
  - **Data Cleansing –** we will do this in this lab
  - **Data Transformation –** not at this time
  - **Data Reduction –** not at this time

- **Once your data are prepared you can mine it using the following process:**



- **Sample** – since your dataset is small enough there is no need to generate a sample
- **Explore** – we did this on assignment 1
- **Modify** – no need to do this for this lab but maybe next week
- **Model** - entails the application of various data mining techniques to develop / adjusting models which can be used to provide useful insight.
- **Assess**
- **Refine and Repeat**

Data Preparation (already accomplished in Assignment 1):

You are part of the management team that has been assigned a store. Your job is to make this store more profitable.

For the purposes of this assignment, a "Data Warehouse" has been created on the Teradata machine. This database has transaction information from all stores for the period of January 1st through 31st 2000.

Data Analysis and Cleaning:
From earlier in the semester (Assignment 1) you should have analyzed your data and performed a variety of statistical analyses on your dataset. You should have also found some dirty data/outliers and told me about them in your summary document. All of you should have different stores with different dirty data but I believe it to be located in the Tender_Amt and Total_Visit_Amt columns.

1) In Teradata SQL Assistant (TSA) you should be able to isolate this bad data by using a query something along the lines of (please keep in mind that your store MAY have valid purchases over $10,000 so you may have to adjust this value accordingly – a valid purchase would be one that makes sense mathematically!):

```
select * from mystore m where m.total_visit_amt > 10000
```

The first few columns of this query should bring you back a result of something like this (please keep in mind that you have different stores so your numbers will be different:

Answerset 1

	Visit_Nbr	Store_Nbr	Register_Nbr	Holder	Membership_Nbr	Member_Code	Tender_Type	Tender_Amt	Sales_Tax_Amt	Total_Visit_Amt	To
1	236690568	3	3	10	12863346	V	8	500000.00	8.25	250000.00	1
2	236690568	3	3	10	12863346	V	8	500000.00	8.25	250000.00	1
3	236690568	3	3	10	12863346	V	8	500000.00	8.25	250000.00	1
4	236690568	3	3	10	12863346	V	8	500000.00	8.25	250000.00	1
5	236690568	3	3	10	12863346	V	8	500000.00	8.25	250000.00	1
6	236690568	3	3	10	12863346	V	8	500000.00	8.25	250000.00	1

… continued…

	Item_Quantity	Total_Scan_Amount	Unit_Cost_Amount	Unit_Retail_Amount	Tax_Collect_Code	Item_Nbr	Category_Nbr	Sub_Category_Nbr	Primary_Desc	Secondary_Desc
8	1.00	22.28	22.1800	22.28	1	3670815	70	18	MAP OF THE WORLD	RETURN APRIL 15
8	1.00	22.88	21.8200	22.88	1	3595779	1	73	GOURMET ASSORTED	TRUFFLE 16 OZ
8	1.00	26.99	25.4700	26.99	1	2902950	31	16	BLACK CARTRIDGE	2 PACK BCI-21
8	1.00	27.50	25.8400	27.50	1	3617871	6	36	KEEPSAKE ALBUM	PHOTO STORAGE
8	1.00	34.98	32.8800	34.98	1	3645878	70	13	FROM SCRATCH	COMPUTER ASSORTMENT
8	1.00	41.93	38.2200	41.93	1	3608994	29	18	I SPY SPOOKY MANSION	PC CDROM

Notice that all 6 of these items were on the same visit number of 236690568.

2) You need to update the Tender_Amt and Total_Visit_Amt columns for this visit to the correct amount so the next step is to determine those correct values. If you add up the values of the Total_Scan_Amount this should be the value of the Total_Visit_Amt. You can add them up manually with a calculator or use the following query:

```
select sum(total_scan_amount)from mystore m where m.visit_Nbr = 236690568
```

Which gave me the value of 176.56

The Tender_Amt = Total_Visit_Amt + Sales_Tax_Amt (176.56+8.25) or you can use the following query to do the same thing!

```
select sum(total_scan_amount)+max(Sales_Tax_Amt) from mystore m where m.visit_Nbr = 236690568
```

Which gave me the value of 184.81

3) The final step is to update/fix your data in your DataMart using an "UPDATE" statement:

```
update mystore set Total_Visit_Amt = 176.56, Tender_Amt = 184.81 WHERE visit_Nbr = 236690568
```

Congratulations!!! Your DataMart is now clean and ready to mine.

Data Mining:

Once the data is prepared, and cleansed then you are ready to perform data mining to discover hidden pieces of information which are actionable. Many different models can be run in parallel and findings can be utilized to cross-reference the validity of the different models developed as a result of mining. We will only perform the Market basket analysis (aka association analysis) and Decision Tree analyses today.

Introduction to Data Mining Utilizing Teradata Warehouse Miner

To begin using Teradata Warehouse Miner (TWM), start the program from the start menu. Create a new project called DataMiningLab1 and perform the following data mining algorithms on your newly cleaned DataMart.

Association Discovery using Teradata Warehouse Miner

What is Association Discovery?

Association discovery is the identification of items that occur together in a given event or record. This technique is also known as market basket analysis. On-line transaction processing systems often provide the data sources for association discovery. Association discovery rules are based on frequency counts of the number of times items occur alone and in combination in the database. They are expressed as "if item A is part of an event, then item B is also part of the event X percent of the time."

The rules should not be interpreted as a direct causation, but as an association between two or more items. Association analysis does **not** create rules about repeating items, such as "if item A is part of an event, then another item A is also part of the event X percent of the time." In association analysis, it doesn't matter whether an individual customer buys one or multiple units of item A: only the presence of item A in the market basket is relevant. However, Identifying creditable associations between two or more different items can help the business technologist make decisions such as when to distribute coupons, when to put a product on sale, or how to present items in store displays.

Listed below are some hypothetical association discovery rules:

- If a customer buys shoes, then 10% of the time he also buys socks.
- A grocery chain may find that 80% of all shoppers will buy a jar of salsa when they also purchase a bag of tortilla chips.
- When "do-it-yourselfers" buy latex paint, they also buy rollers 85% of the time.
- Forty percent of investors holding an equity index fund will have a growth fund in their portfolio.

These example rules have a left-hand side (antecedent) and a right-hand side (consequent). For example, for the first rule listed above, shoes is the antecedent item and socks is the consequent item. Both sides of an association rule can contain more than one item.

Before we can continue with our association discovery we need to prepare and designate the data which will be used.

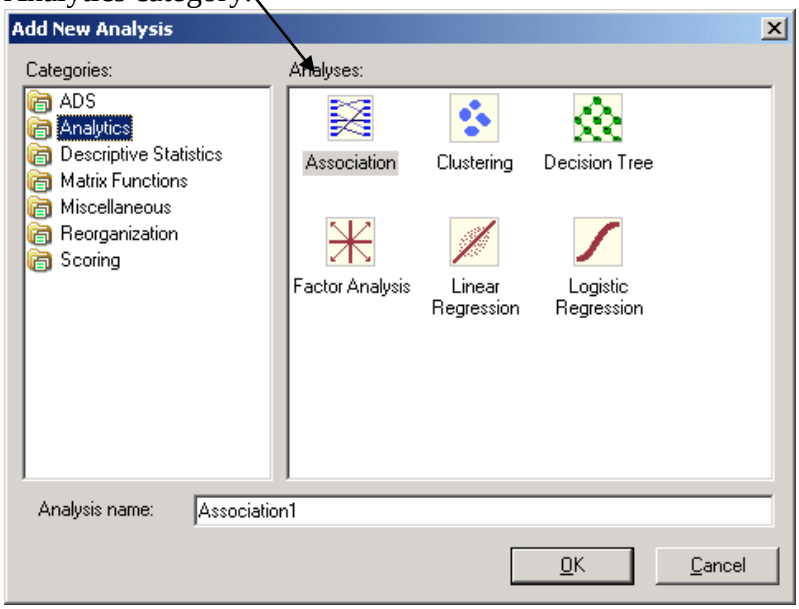
Data Preparation:

To perform association discovery, the input data set must have a separate observation for each product purchased by each customer, as illustrated in the following example.

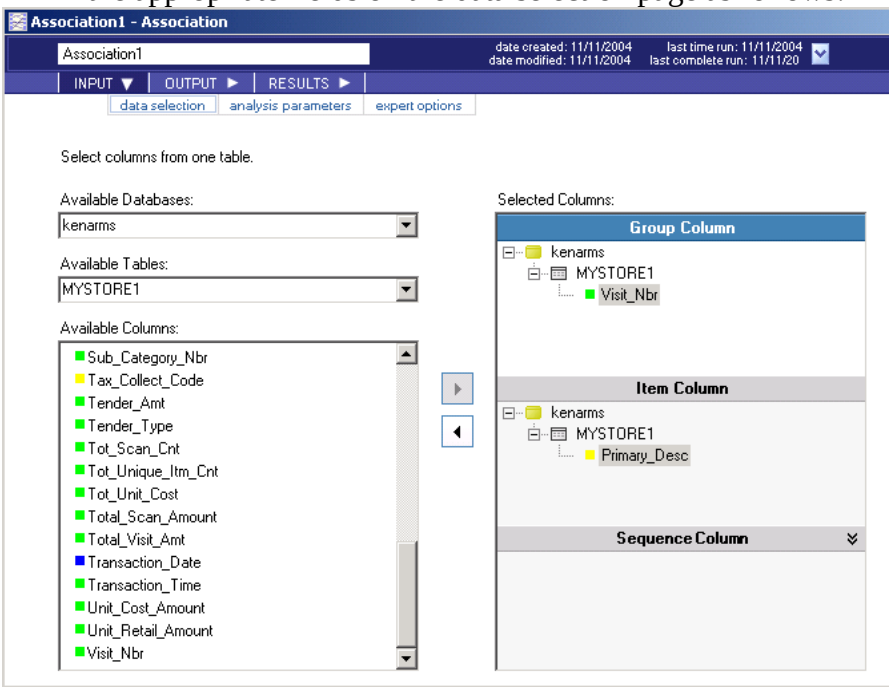
Customer	**Product**
Ken Armstrong visits the store on 1/5/2000	Natural Choice Tomato Soup
	Washington Carver Honey Ale
Mary Lamb visits the store on 1/6/2000	Natural Choice Black Bean Soup
	Otto's Oatmeal

Thus, we would need to define the DataMart called MyStore we have created earlier as our input source. A consideration at this stage would be whether you want to include the whole data set or a subset of the data (by defining the sample size) in running the association. Since there are so few sample (yes, 150,000 is just a few ☺) there is no need to split your input dataset into pieces for a market basket analysis.

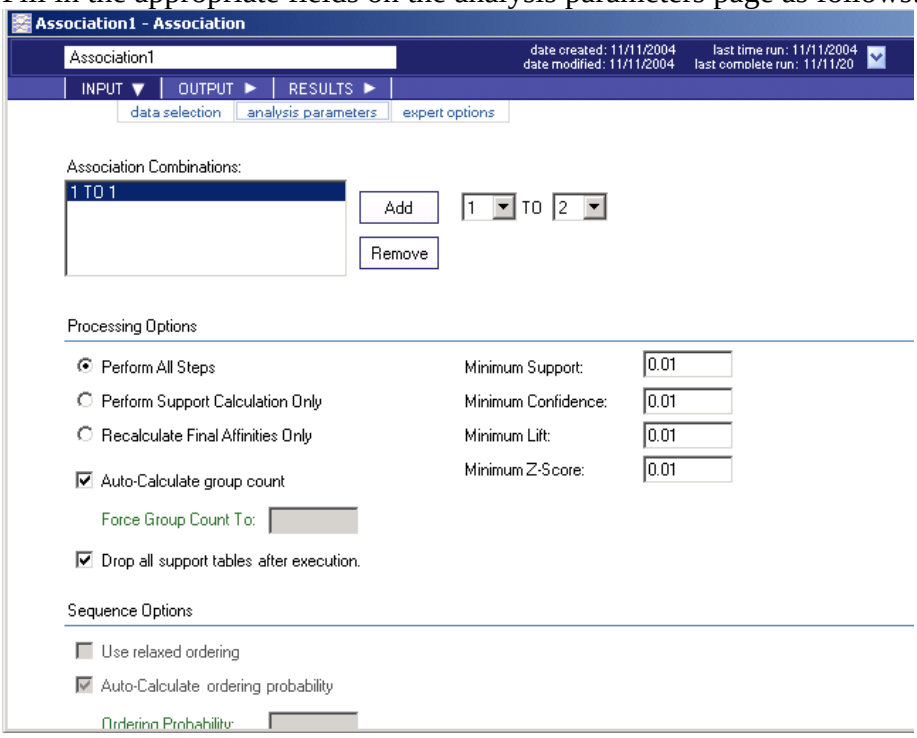
Create a new project and call it Data Mining Lab 1:
In order to perform an association, add a new analysis and select Association from the Analytics category.



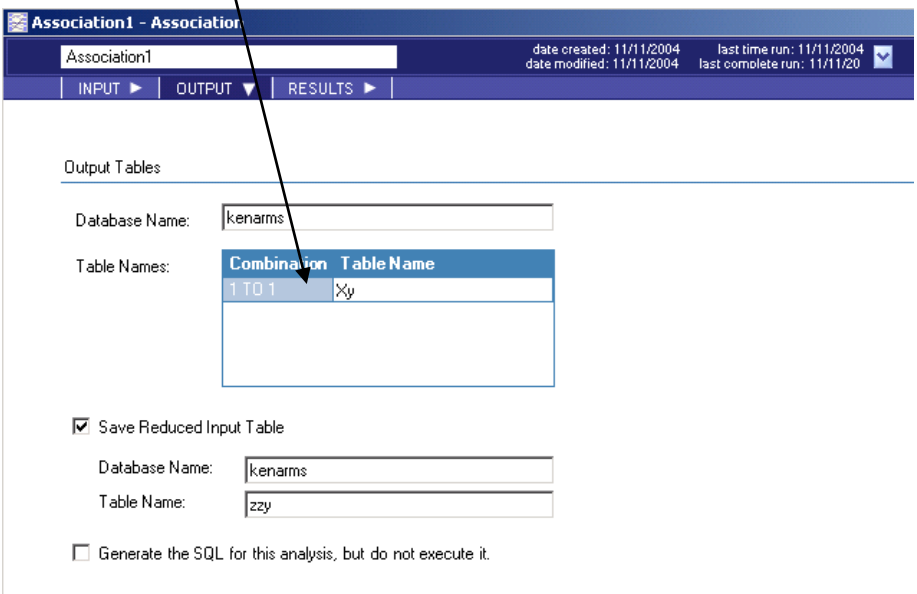
Fill in the appropriate fields on the data selection page as follows:



Fill in the appropriate fields on the analysis parameters page as follows:

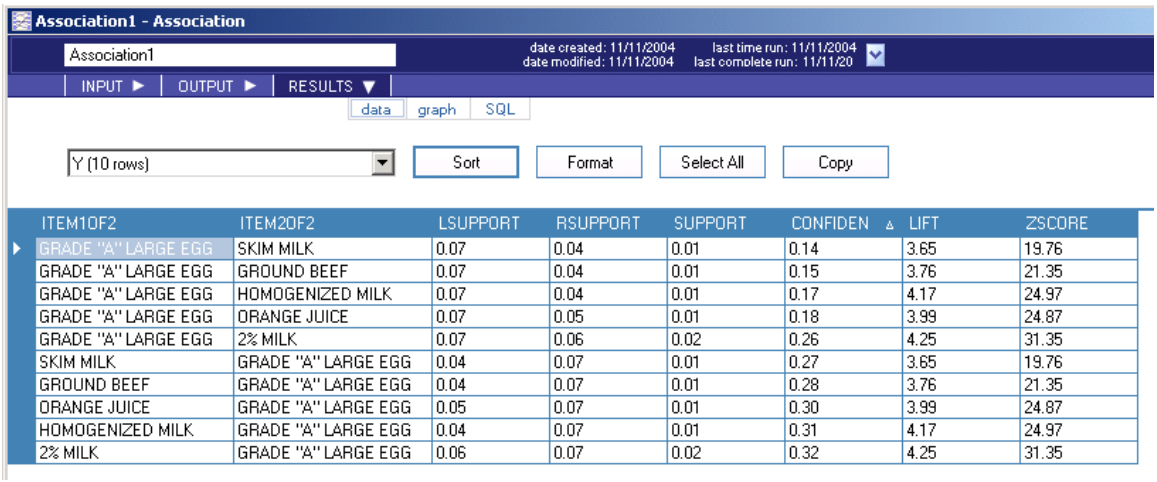


If you want to store your output into a table in your ES842xx database for later analysis/exporting to Excel/Reporting, etc., click on the output screen and put in a table name that will be required to hold the results of your analysis.



If you want to export this data, you can then go to SQL Assistant and run a query there like: SELECT * FROM Xy and then select a cell in the answer set window then click on file/copy to notepad. You can then copy/paste the data into Excel on your local machine if you like.

Run your analysis and look at your results. Adjust your minimum threshold values as required to get more/less data. You should get something along these lines.

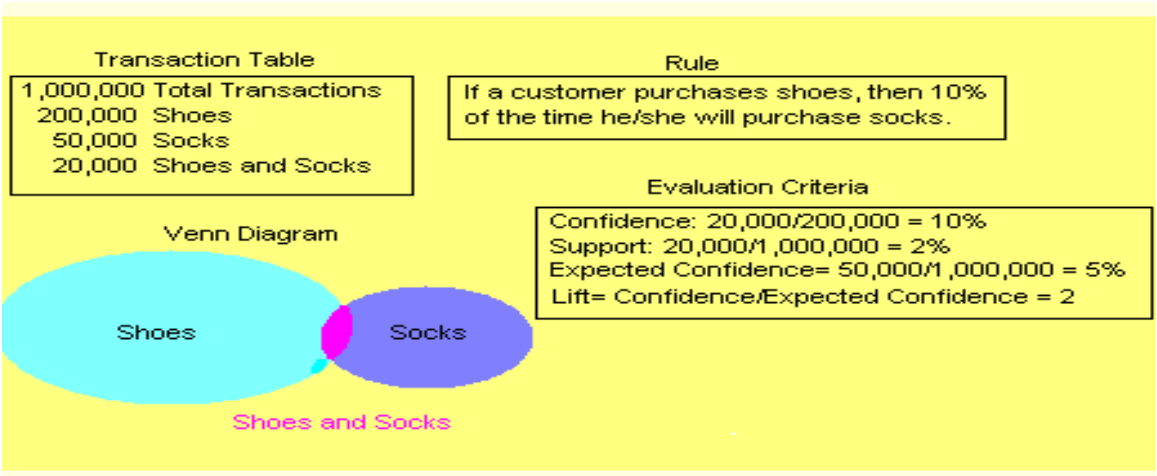


ITEM1OF2	ITEM2OF2	LSUPPORT	RSUPPORT	SUPPORT	CONFIDEN	LIFT	ZSCORE
GRADE "A" LARGE EGG	SKIM MILK	0.07	0.04	0.01	0.14	3.65	19.76
GRADE "A" LARGE EGG	GROUND BEEF	0.07	0.04	0.01	0.15	3.76	21.35
GRADE "A" LARGE EGG	HOMOGENIZED MILK	0.07	0.04	0.01	0.17	4.17	24.97
GRADE "A" LARGE EGG	ORANGE JUICE	0.07	0.05	0.01	0.18	3.99	24.87
GRADE "A" LARGE EGG	2% MILK	0.07	0.06	0.02	0.26	4.25	31.35
SKIM MILK	GRADE "A" LARGE EGG	0.04	0.07	0.01	0.27	3.65	19.76
GROUND BEEF	GRADE "A" LARGE EGG	0.04	0.07	0.01	0.28	3.76	21.35
ORANGE JUICE	GRADE "A" LARGE EGG	0.05	0.07	0.01	0.30	3.99	24.87
HOMOGENIZED MILK	GRADE "A" LARGE EGG	0.04	0.07	0.01	0.31	4.17	24.97
2% MILK	GRADE "A" LARGE EGG	0.06	0.07	0.02	0.32	4.25	31.35

The lift, confidence factor, and level of support are three important evaluation criteria of association discovery. Lift is the ratio of the confidence factor to the expected confidence. Lift is a factor by which the likelihood of consequent increases given an antecedent. Values of lift greater than one are desirable.

The strength of an association is defined by its confidence factor, which is the percentage of cases in which a consequent appears given that the antecedent has occurred. The level of support is how frequently the combination occurs in the market basket (database). *The total number of transactions in the data set (sample size) should be considered in determining acceptable levels of support.*

The diagram below demonstrates these evaluation criteria.



Note: Information is provided by TWM on the association that relates to lift, confidence, relations, support, and transaction count. Lift is the ratio of the confidence

factor to the expected confidence. The confidence level explains the percentage of cases in which the consequent occurs given the occurrence of the antecedent.

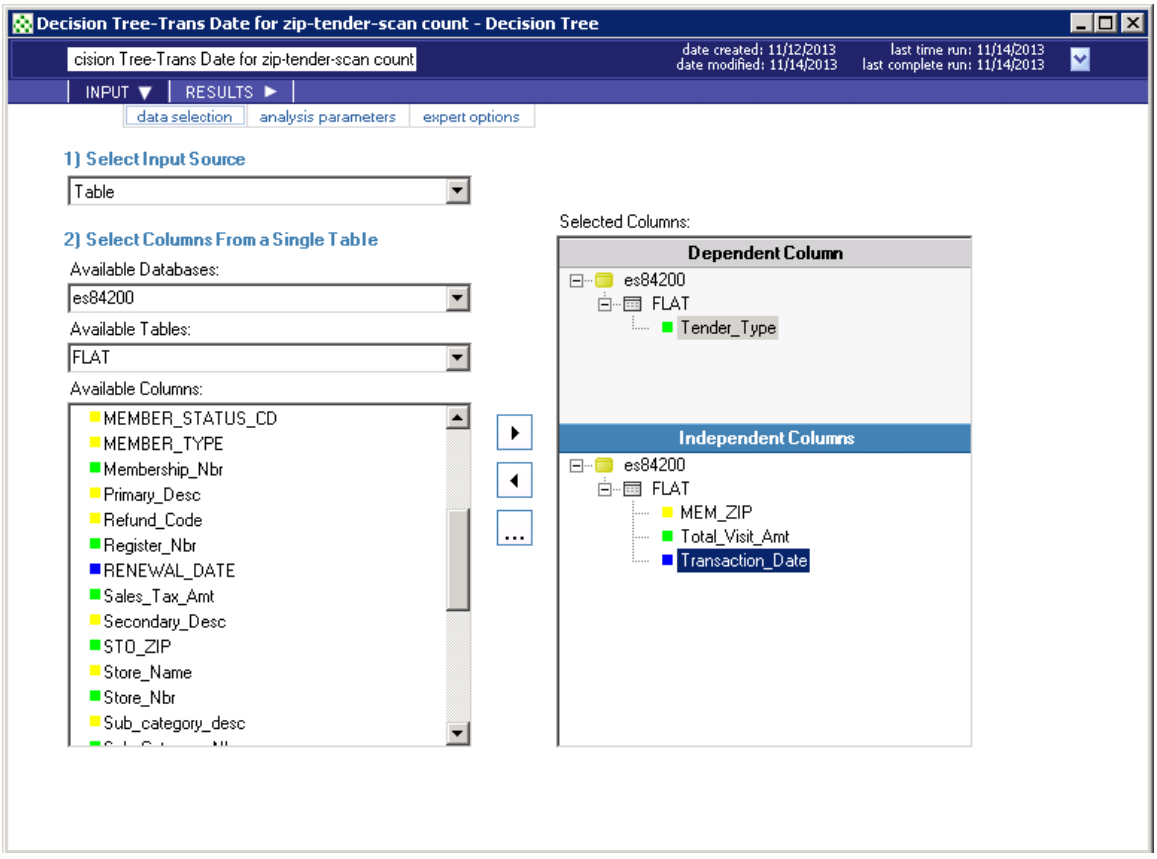
The confidence factor with shoes on the left-hand side and socks on the right-hand side is 10% (20,000/200,000). The lift value of two implies that you are twice as likely to buy socks if you bought shoes, than if you did not buy shoes.

A creditable rule has a large confidence factor, a large level of support, and a value of lift greater than 1. Rules having a high level of confidence but little support should be interpreted with caution.

Decision Tree:

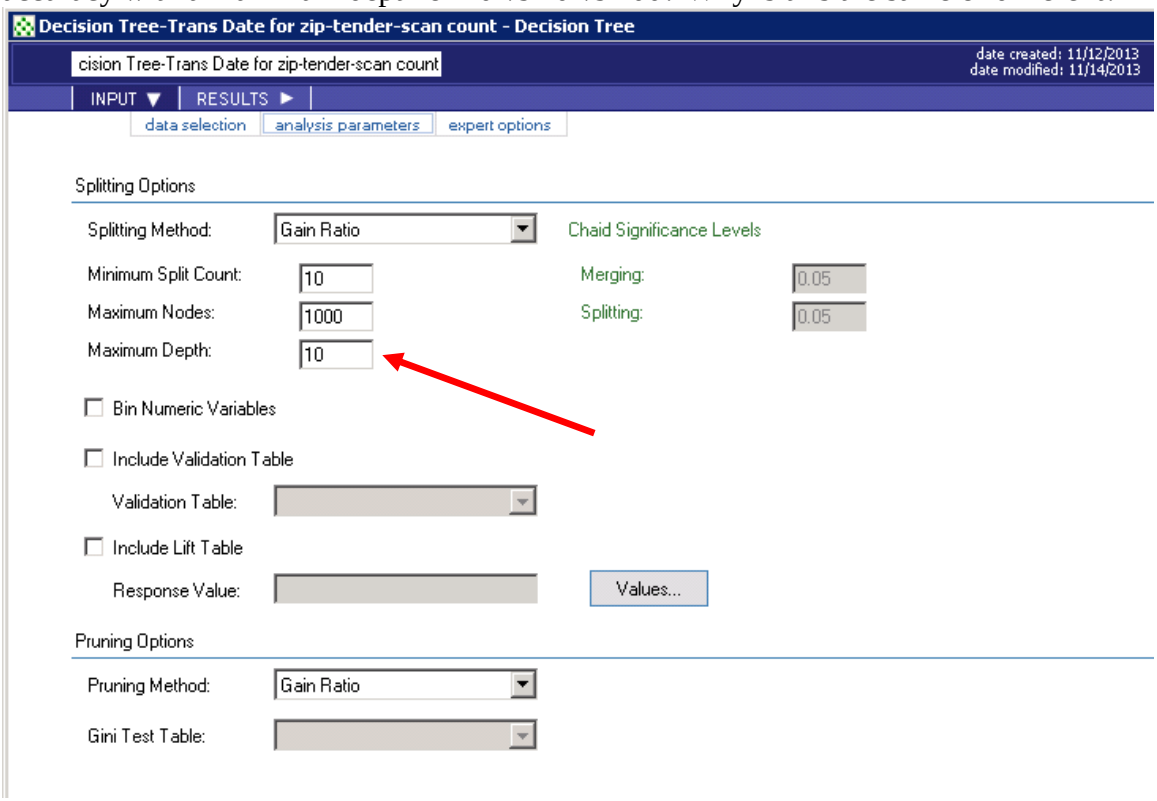
Management wants you to find out if the Total_Visit_Amount affects the likelihood of how the member will pay for their visit (Tender_Type). They also want to know if the date of the month or the member's zip code have any influence on this scenario.

Create a decision tree analysis that looks like the following:



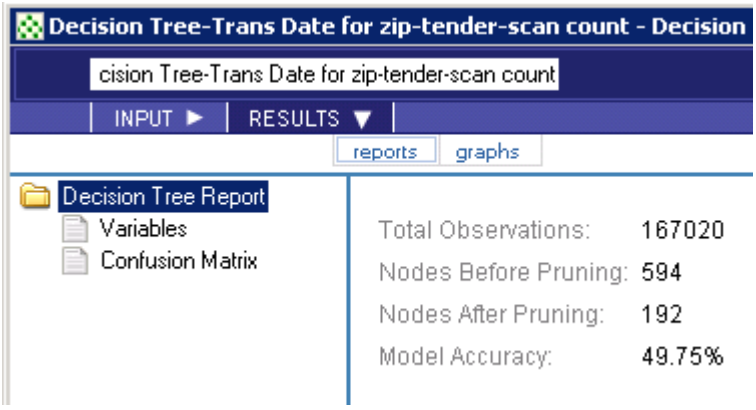
Set the analysis parameters to the following values but do not be afraid to play with other values (like the maximum depth) to see how the model reacts! What is your overall

accuracy with a maximum depth of 10 vs 20 vs 100? Why is this the same or different?



Run the analysis

Under results/reports/Decision Tree Report, analyze your model to determine the overall accuracy.



Under results/reports/variables look to see if all of your original independent variables are still there. Did any of them "fall out"? If so, they were deemed irrelevant to the analysis. What happens to your model accuracy and the look of the tree if you go back to the

input/analysis parameters and vary the maximum tree depth of 5 vs 10 vs 20 vs 100? Why is this the same or different

Next look at your graph (wander through the tree) and look at all of the green (leaves) to determine your probabilities of total visit amount and how it determines the payment type. Which leaf or leaves have the highest likelihood of predicting the Tender_Type used on the transaction?