Example Market Basket Analysis on OE operational database.

In order to perfume data mining in the OE database, multiple steps were taken. The stages that were gone through to do the mining are illustrated below.

The first step was preparing the dataset to be feed into the modelling tool which in this case was Weka. This step involved running a query and fetching the order and product data from the order items table of OE database. The query returns 105 orders and 185 products.

```
SELECT order_id, 1 , product_id
FROM oe.order_items i
ORDER BY order_id, product_id
```

Execute | Save Script | Clear Screen | Cancel

ORDER_ID	1	PRODUCT_ID
2354	1	3106
2354	1	3114
2354	1	3123
2354	1	3129
2354	1	3139
2354	1	3143
2354	1	3150
2354	1	3163
2354	1	3165
2354	1	3167
2354	1	3170
2354	1	3176
2354	1	3182
2356	1	2289

Figure 1

The second step involved transferring the data to Excel and placing them into a pivot table.

SUM OF 1 PRODUCT_ID																		
Order_ID	1781	1782	1787	1791	1797	1799	1803	1806	1808	1820	1822	1825	1910	1912	1948	2058	2093	2211
2358																		
2391																		
2391																		
2439																		1
2391		1			1		1		1									
2358																		
2391																1	1	
2358																		
2439																		
2358																		
2391													1		1			
2391																		
2373																		
2439																		
2439																		
2391																		
2373													1					
2364																		
2451																		
2370										1		1						
2390																		
2433																		
2390																		
2451																		

Figure 2

	1781	1782	1787	1791	1797	1799	1803	1806	1808	1820	1822	1825	1910	1912	1948	2058	2093	2211	2236
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	Yes	?	?
Yes	Yes	?	?	Yes	?	Yes	?	Yes	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	Yes	Yes	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	Yes	?	Yes	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	Yes	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	Yes	?	Yes	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

Figure 3

The data in Figure 2 is not readable by data mining tool, Weka and for this reason; the next step involves making the data readable by the data mining tool by replacing the 1s with Yes and the empty spaces with '?' as Weka uses the '?' symbol as a null value. The file is saved in .csv (comma separated) format. This makes the data readable by the Weka Tool. One particular issue encountered was the large data set that created problem during changing the data to be readable by Weka.

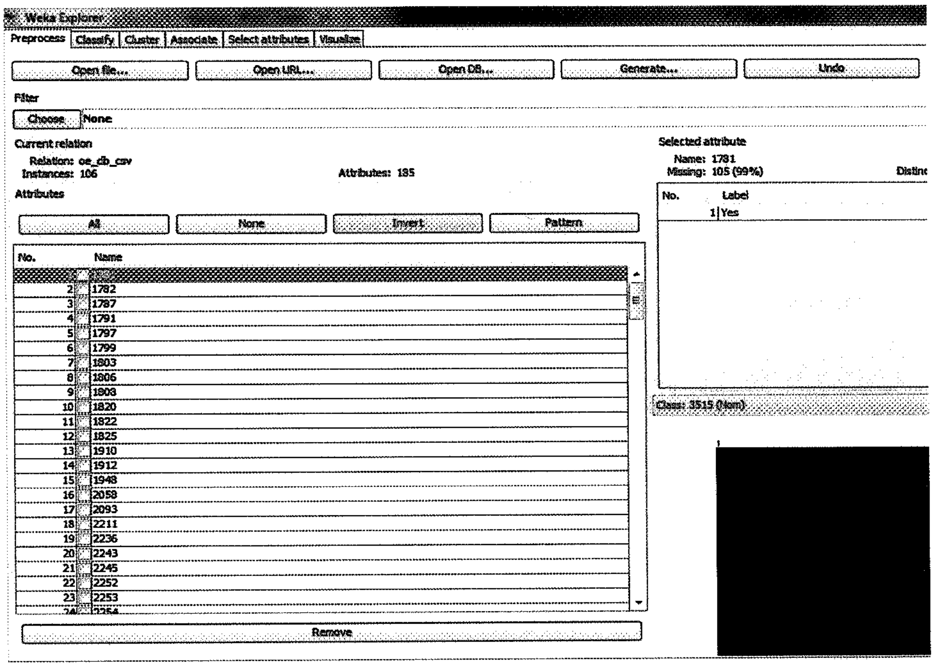


Figure 4

The fourth figure shows the Weka tool after the csv file is loaded. It lists the number of attributes and instances. The file is ready for the association rule to be applied.

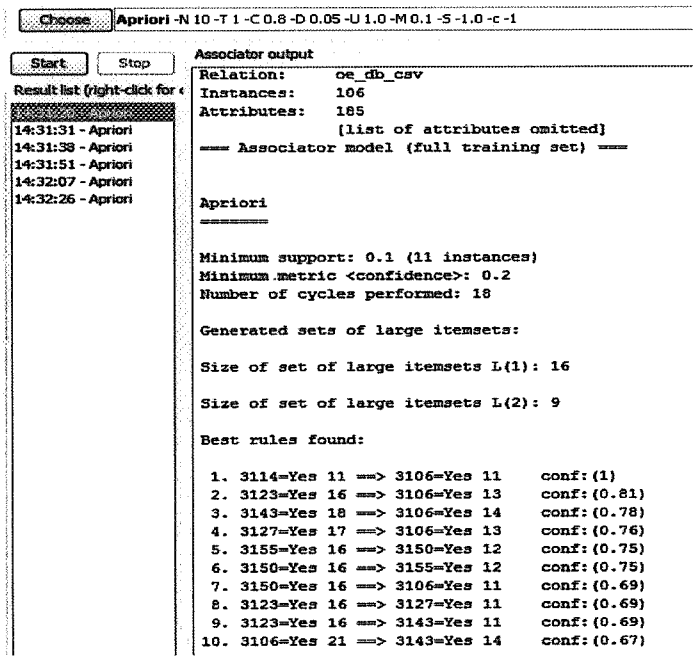


Figure 5

The above figure shows a model that selects all the 185 attributes and assoicates them using "Apriori" Associator algorithm with minimum confidence level of 0.2 or 20%. This gives 10 best rules with different confidence level.

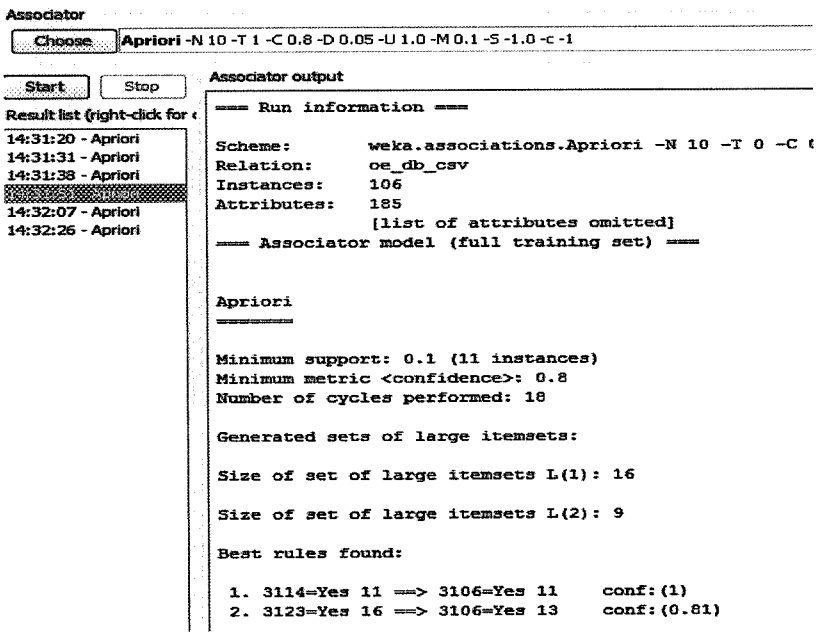


Figure 6

The sixth figure shows the model with 80% and above confidence and we can see that the best rules are reduced to just two from 10 when it was 20% confidence.

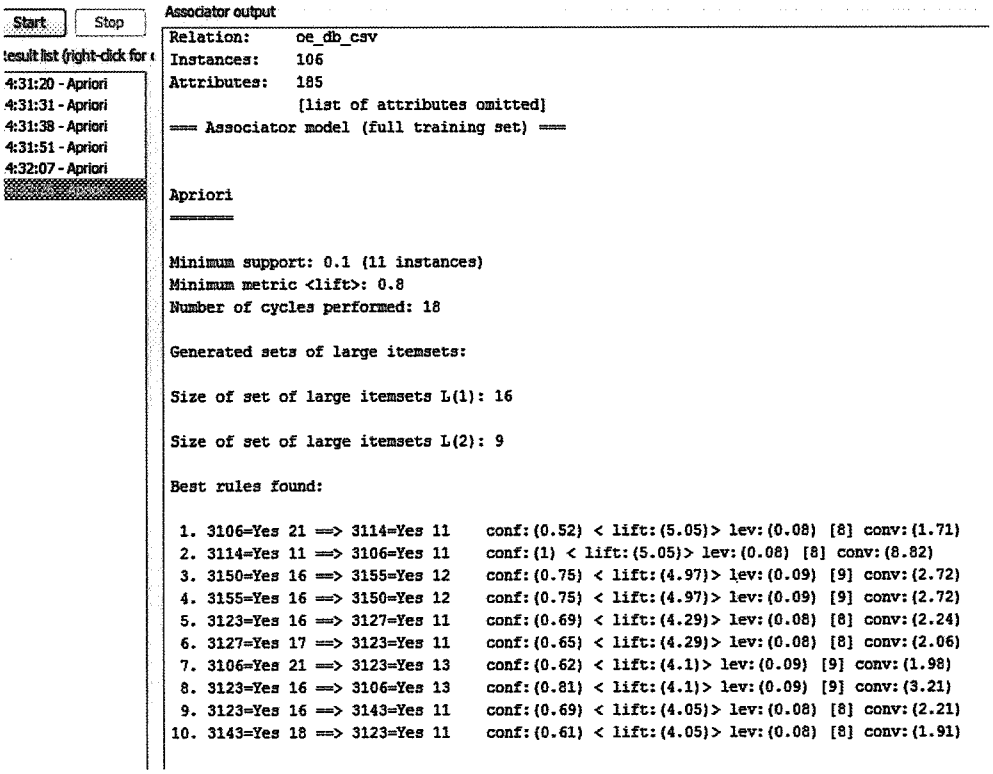


Figure 7

The seventh figure shows the best rules at above 0.8 lift along with the confidence, leverage and conviction of the rules.

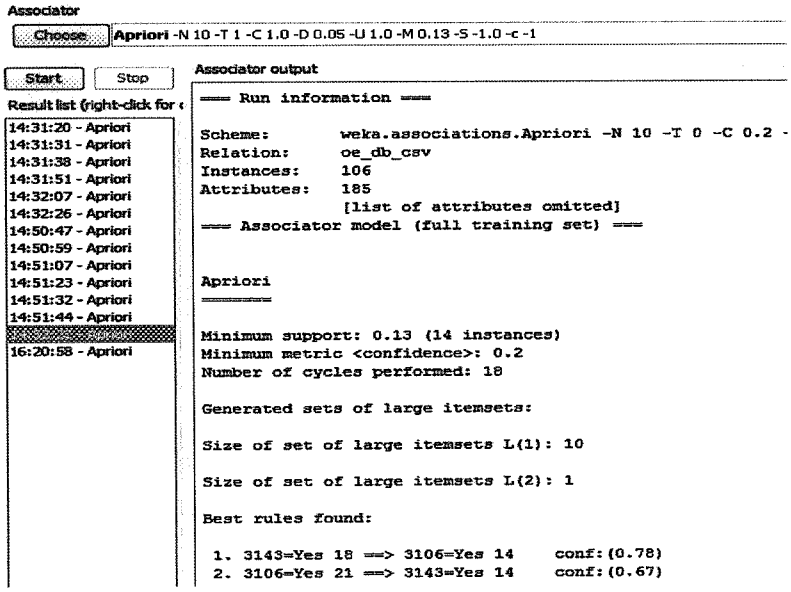


Figure 8

In the eighth figure, the support is changed to .13 from .1 at 20% confidence reducing the best rules to mere 2 sets.

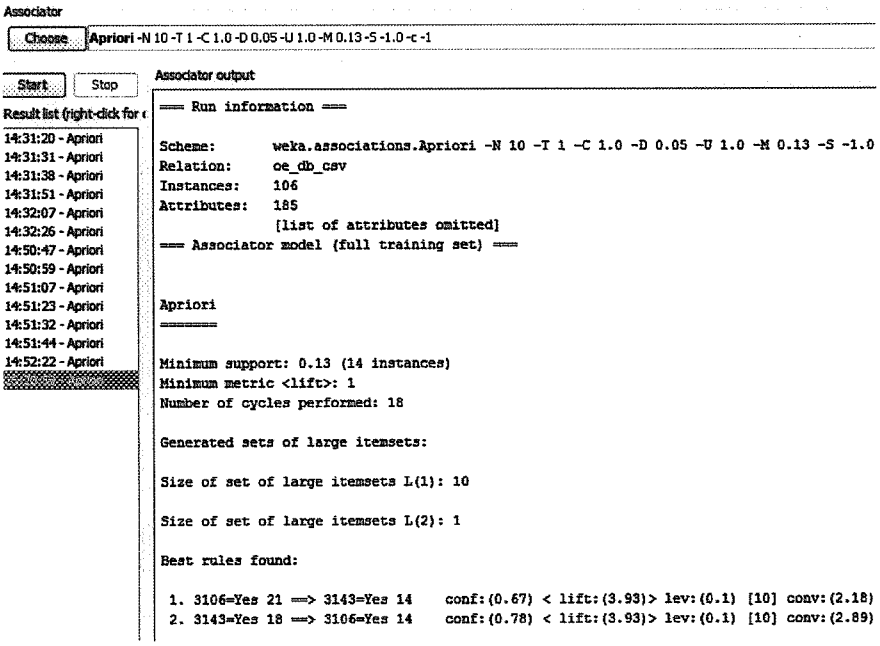


Figure 9

In this figure the support is maintained at .13 while the lift is chosen as the metric type with minimum value of 1 giving the two sets of result with their confidence, lift, leverage and conviction value.

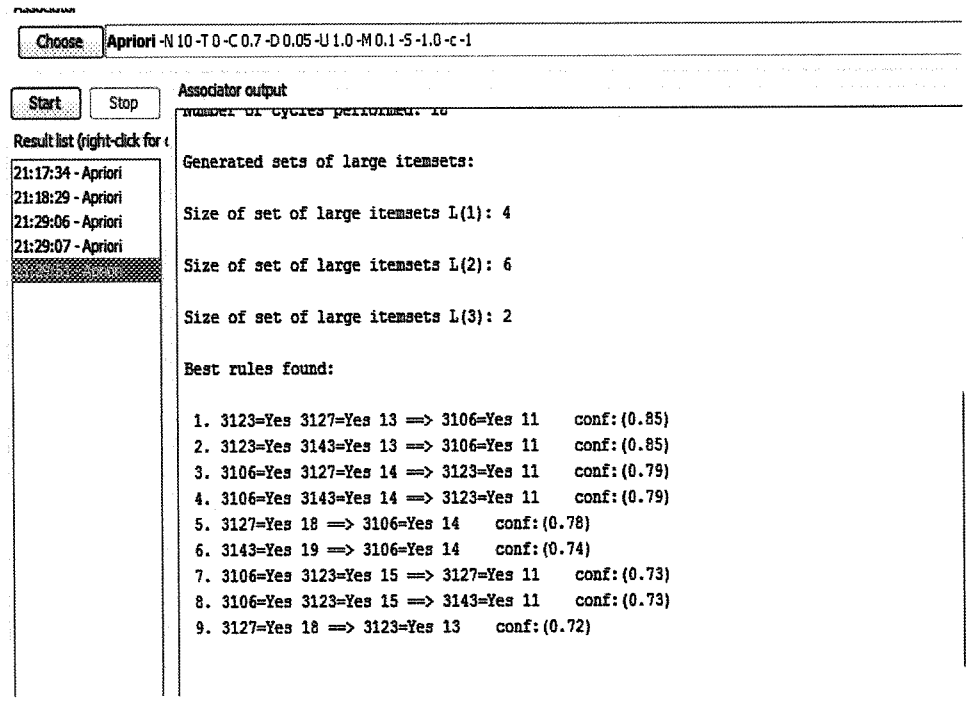


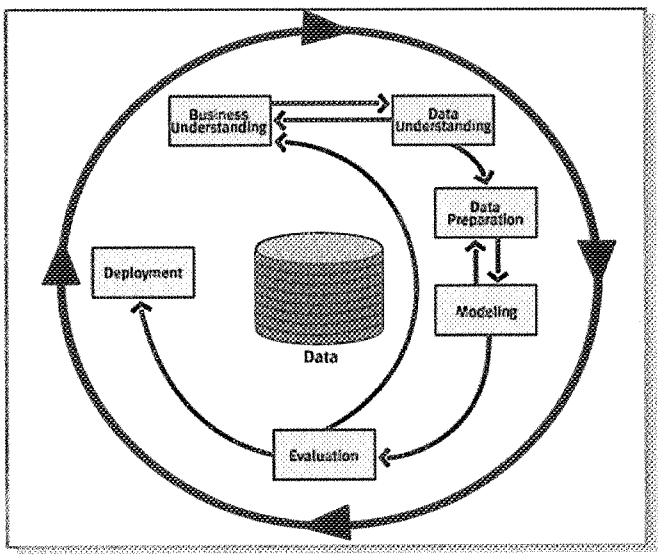
Figure 10

This tenth figure shows best rules for three items with 0.7 confidence an slight modification of the data with the help of the viewer inbuilt in Weka.

Evaluation Report

Decision support systems are a specialised computer systems used to inform and support during the decision making by a management of a business (Webopedia, 2010). This is done by extracting valid, previously unknown information from a large database which are filtered, optimized and organised. This process of trying to find patterns or correlations in the data contained in relational databases is known as data mining. One of the models of data mining is the Market Basket Analysis which is described by Berry & Linoff (2004, p.289) as a technique to understand the customer behaviour through understanding point of sale-transaction-data by using different rules.

The data mining project was based on the CRISP (Cross-Industry Standard Process for data mining) methodology which involves the six phases in its life cycle as shown in the diagram (CRISP-DM Consortium, 2000).



CRISP-DM Life-cycle

The data mining project started for improving the profit of business by establishing a method of analysing the data which could enable the business to target a wider audience. One purpose was to identify the products relationship during sale and the data was gathered using iSQL *Plus from the OE database as we can see in Figure 1. The raw data is then cleaned and transformed in Microsoft Excel using Pivot feature as shown in figure 2 and 3 which is then feed into the modelling tool as seen in figure 4. The modelling tool used during this project was WEKA 3.6 and different association metric were applied to build different models.

During the modelling process, the association rule was applied by the tool which implements the a-priori algorithm. Berry & Linoff (2004, p.296) states that the results from the association rules could be categorised into three general rules: actionable, trivial and inexplicable. The actionable rules contain information that are useful and can be acted upon. In our mining process, the rules like 'if motherboard (3114) then keyboard (3106)' or if 'power supply (3123) then keyboard (3106)' from figure seven could be considered actionable rules as they have high confidence and good lift. The

rules like 'if printer (3127) then power supply (3123)' or 'if power supply (3123) then screws (3143)' from figure five could be considered a trivial rule as this would be familiar knowledge for the people in that industry. Inexplicable rule have no explanation and does not suggest a course of action that could be implemented which is the case with the rule 'if a card holder (3150) then keyboard (3106)' rule in figure five. When there is a large amount of data, the analysis of all items can't be done on individual basis and to overcome this Berry & Linoff (p 305-306) suggests using product hierarchies by using more general items initially and to repeat the rule while focusing on specific items and their transactions. In the case of the analysis carried out for the OE application, there is a product category Id but the database lacks a distinct field for the category name barring us from creating a hierarchical taxonomy of the items.

Market basket analysis of the data at three levels: order, items and customers provides us a remarkable way to know the customer and comprehend their different purchase behaviour (Gutierrez, 2008). Association technique provides us with different rules to find products which tend to sell together by using three measures: support, confidence and lift (Berry & Linoff, p.319).

There are different methodologies like CRISP-DM and SEMMA for data mining but it is not fully adopted throughout the industry (KDnuggets, 2007). The tools like Excel and iSQL *Plus are widely used and well documented but even though WEKA is a stable tool, it requires highly technical knowledge of the subject area to use it in a productive way. The steps performed during this data mining project did not pose a huge challenge to do a high level data mining with market basket analysis model with association techniques to find some association on the OE operational database. But in order to provide highly actionable models, the data mining operation needs to be performed on an OLAP system and will require a highly qualified group of people. In conclusion, if the human resources as well as system resource can be overcome, mining data to provide decision support can produce highly profitable results to the company.

References:

Webopeida, 2010. *Decision support system* [online]. Available: http://www.webopedia.com/TERM/D/decision_support_system.html [accessed 01 March 2010]

GUTIERREZ, N., 2006. *Demystifying Market Basket Analysis* [online]. Available: http://www.information-management.com/specialreports/20061031/1067598-1.html?pg=1 [accessed 02 March 2010]

BERRY, M.J.A. and G. LINOFF, 2004. *Data Mining Techniques*. 2nd ed. London: Wiley

KDNUGGETS, 2007. *Data Mining Methodology* (Aug 2007) [online]. Available: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm [accessed 02 March 2010]

CRISP-DM CONSORTIUM, 2000. *CRISP-DM 1.0: Step-by-step data mining guide* [online]. Available: http://www.crisp-dm.org/CRISPWP-0800.pdf [accessed 02 March 2010]