Engaging Business Students with Data Mining

Dan Brandon
Professor, MIS Department
Christian Brothers University
Memphis, TN

ABSTRACT

The Economist calls it "a golden vein", and many business experts now say it is the new science of winning. Business and technologists have many names for this new science, "business intelligence" (BI), "data analytics," and "data mining" are among the most common.

The job market for people skilled in this area is growing rapidly. ComputerWorld's Survey of its 100 IT leaders ranked it as their top file priority for 2014, and a Gartner survey of 1,400 chief information officers suggests that business intelligence is the number one technology priority for IT organizations.

For these reasons, colleges are rushing to develop curriculums, courses, and teaching methods to prepare students for this field. Teaching business students this new science is challenging for a number of reasons including the fact that it uses a variety of disciplines, many traditionally outside of the business school including sophisticated computer algorithms. Thus "engaging" business students with lessons about data mining can be challenging. In this paper, a method of such teaching engagement is discussed and illustrated.

BACKGROUND

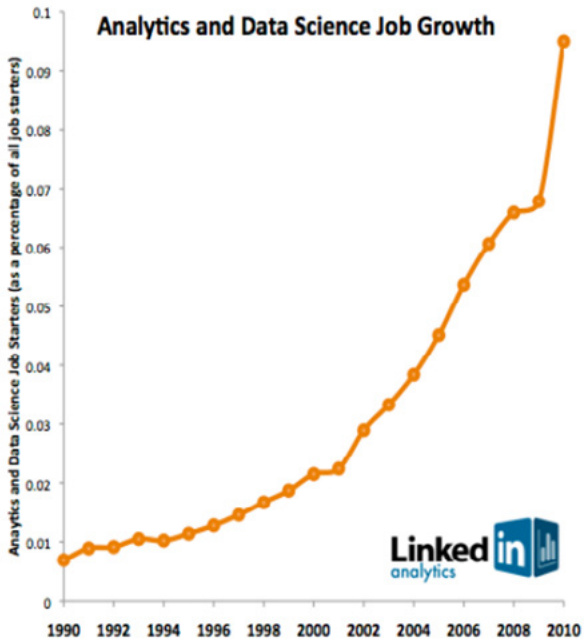
In earlier times businesses had a close physical relationship with their customers and had much first-hand knowledge about their customers such as whom they were, were they lived, what were their needs, and so on. However, as businesses became larger and more global in scope, it became harder for them to understand who their customers are, how to best serve them, and how to maximize their own profits.

To make such decisions in today's fast paced global marketplace, companies make extensive use of something called "business intelligence". This approach relies on large data warehouses and complex computer algorithms to sift through endless amounts of data. Business technologists have many names for this revolutionary technology; "business intelligence" (BI), "data analytics," and "data mining" are among the most common

The Economist says it's "a golden vein", and many business experts now call it "the new science of winning". It's been adopted by nearly every Fortune 500 company. Even many professional sports franchises are using use this new technology. A Gartner survey of 1,400 chief information officers suggests that business intelligence is the number one technology priority for IT organizations".

Most companies are not short on data. Large businesses store hundreds of terabytes just from their daily transactions. This tells them who is buying what, and also where and when. But today business also needs to know why, or why not.

Traditionally this was done with classical business research such as surveys, focus groups, etc. But today it also comes from web and social media such as tweets, videos, likes, and "clickstream data". This is typically called "Big



Data". The average large company now has more data stored than the Library of Congress.

Job growth in this area is very strong as illustrated in the figure below. InformationWeek's 2012 State of IT Staffing Survey reveals that 40% of those employers who cite big data and analytics as a top hiring priority say they'll increase staffing in these areas by 11% or more during the next two years. At the same time, 53% of these companies say it will be hard to find big-data-savvy analytics experts. A Gartner survey of 1,400 chief information officers suggests that business intelligence is the number one technology priority for IT organizations.

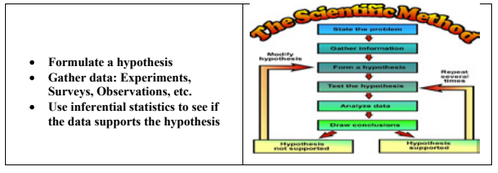
ComputerWorld's Survey of its 100 IT leaders ranked their top five priorities for 2014:

- Business intelligence
- Mobility (tablets, apps, etc)
- Application development
- Cloud computing
- Security

BUSINESS INTELLIGENCE AND DATA MINING

Wikipedia defines business intelligence (BI) is a set of theories, methodologies, architectures, and technologies that transform raw data into meaningful and useful information for business purposes.

Business intelligence, particularly via data mining, reverses the traditional "scientific method" which has these sequential steps:



Wikipedia defines data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. With data mining, one starts with the data and then applies mathematical and computational algorithms to determine the supported hypotheses. Essentially one is "mining" for hypotheses or rules. There are a number of applications of data mining including:

- Association or affinity analysis – looking for statistical rules among data items
- Nearest-neighbor and clustering method – looking for concentrations of data in n-dimensional space
- Text mining and context analysis–deriving quality information and patterns from text

AFFINITY ANALYSIS

Perhaps most used and most successful of the data mining applications is affinity analysis which uses a specialized set of algorithms that sort through large data sets and expresses statistical rules among items. A typical usage is for analyzing purchase patterns of customers via transaction data which contain a huge wealth of information that can be used for a variety of purposes as:

- Marketing
- Up selling
- Cross selling
- Recommendations
- Inventory & logistics
- Product placement
- Store management

Affinity analysis is also called "market basket analysis" since it essentially determines what products people purchase together. Stores can use this information to place these products in the same area (particularly preferred brands), direct marketers can use this information to determine which new products to offer to their current customers, and inventory policies can be improved if reorder points reflect the demand for the complementary product.

Affinity analysis finds rules which are derived in the form "left-hand side implies right-hand side". An example is:

Yellow Peppers IMPLIES Red Peppers, Bananas

The rules are unidirectional, and the following is an "obvious" rule:

Caviar IMPLIES Vodka

But the reverse is not true:

Vodka IMPLIES Caviar

The key measures of mining predictive ability of a rule are:

- Support (prevalence) refers to the percentage of baskets where both left and right side products were present
- Confidence measures what percentage of baskets that contained the left-hand product also contained the right
- Lift measures how much more frequently the left-hand item is found with the right than pure chance (the product of their individual probabilities of occurrence)

For affinity analysis, we first need a list of transactions of what was purchased–this is readily available with modern electronic cash registers. A transaction is the purchase of one or more items by a customer at one point in time and space – a "shopping cart" or "market basket".

Next, we choose a list of products to analyze (perhaps all our products), and tabulate in a table how many times each was purchased with the others. The diagonals of the table shows how often a product is purchased in any combination, and the off-diagonals show which combinations were bought.

Consider the following simple example of five transactions at a convenience store:

- Transaction 1: Frozen pizza, cola, milk
- Transaction 2: Milk, potato chips
- Transaction 3: Cola, frozen pizza
- Transaction 4: Milk, pretzels
- Transaction 5: Cola, pretzels

Below is the resulting "affinity table". We notice that Pizza and Cola sell together more often than any other combo (perhaps a cross-marketing opportunity), and also notice that Milk sells well with everything (perhaps people probably come here specifically to buy it).

Product Bought	Also bought:				
	Pizza	Milk	Cola	Chips	Pretzels
Pizza	2	1	2	0	0
Milk	1	3	1	1	1
Cola	2	1	3	0	1
Chips	0	1	0	1	0
Pretzel	0	1	1	0	2

From the affinity table we want to find association rules which suggest a relationship between items in the transaction. The rules are written as for single items A and B:

A IMPLIES B (or A → B)

Support is calculated as the % of transactions (baskets) where an association rule applies, that is where we see both item A and B in the same basket. For example, if 500 baskets contain both A and B out of a total of 1000 baskets, then the support is 50%. A implies B and B implies A both have the same support. The support measure for Cola IMPLIES Pizza is 40% (2/5); of the 5 transactions 2 have both cola and pizza. Note support does not consider direction (Pizza IMPLIES Cola is also 40%).

Confidence measures the predictive accuracy of a rule, and it is defined as the probability that item B is in the basket if item A is in the basket ("conditional probability")

$$\mathbf{P(B|A) = P(AB)/P(A)}$$

It is calculated as the support (A & B)/P(A) where support (A) is the % of baskets containing A. For example, if 500 baskets contain both A and B out of a total of 1000 baskets, then the support of A & B is 50%. If A is in 75% of baskets, the confidence is 50/75 or 67%. Milk IMPLIES Chips has a confidence of 33%, since the support of "Milk plus Chips" is 20% (1/5) and Milk is in 60% of baskets (3/5). Thus 20%/60% is 33. Confidence is unidirectional.

Lift is calculated as the ratio of support to a product to the individual probabilities of both sides:

$$\mathbf{P(AB)/(P(A) * P(B))}$$

For example, if 500 baskets contain both A and B out of a total of 1000 baskets, then the support of A & B is 50%. If A is in 75% of baskets and B is in 20% of the baskets, then the lift is: .50/(.75*.20) = 3.33. Computing Lift: TABLE: Lift is the ratio of support of a product to the individual joint probabilities of both sides. Cola IMPLIES Pizza lift is .40/(.60 * .40) = 1.67.

The rules can be formulated for each pair of products, and the three measures calculated. Only the rules that have significant measures are going to be accepted – this is the mining portion of the process. Some rules are going to be trivial (hot dogs and buns sell together), and some rules may be far from obvious.

Engaging Students via Interactive Web Teaching Tools

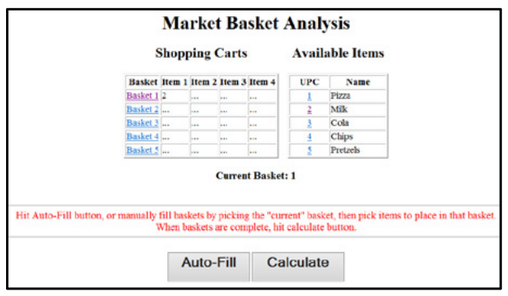
Due to the business need for data mining and the resulting strong job market, colleges are rushing to develop curriculums, courses, and teaching methods to prepare students for this field. It is a field that requires both understanding of the business need and application of data mining but also the underlying technology. Thus teaching business students this new science is challenging for a number of reasons including the fact that it uses a variety of disciplines, many traditionally outside of the business school including database design, programming, and sophisticated computer algorithms.

Our teaching approach is to first describe the data analytics method and its business purpose. Next the student is provided with a basic interactive and intuitive tool that "engages" him. The tool is programmed in HTML5 and JavaScript. The engagement tool for affinity analysis al-

lows the student to interactively fill shopping baskets with available items. The first screen shot below shows the starting screen where the student can manually place items into the baskets, or hit the "auto-fill" button to fill the carts.

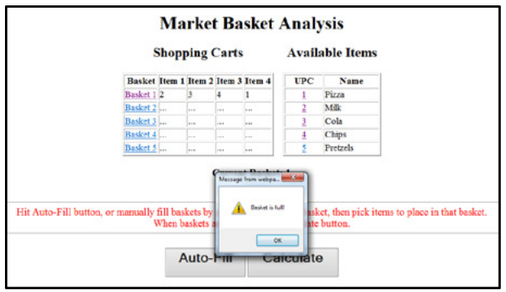
The screen shot below shows the screen after basket selection. The basket is highlighted and the "current basket" number is set.

The next screen shot below shows the placing of item2 in basket 1, simply by clicking on the UPC code.

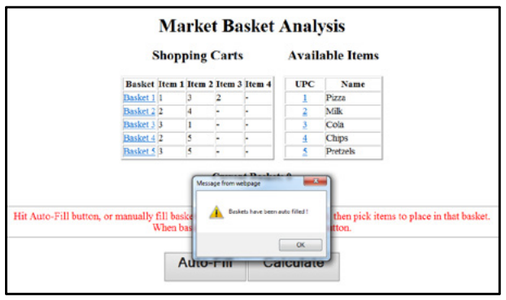


Placing item 2 in basket 1, when item 2 is already in that basket produces the error shown in the screen below.

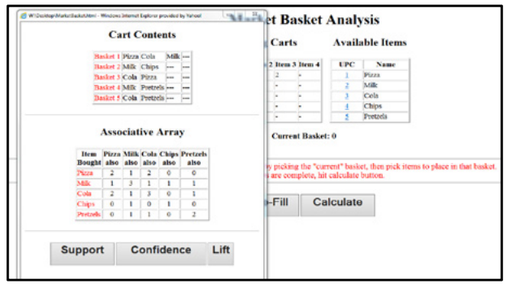
Similarly, exceeding the capacity of a basket gives the error shown in the screen shot below.



As well as manually selecting baskets and items (with the mouse), there is an "auto-fill" option to fill the baskets with items matching the "example case" previous described. The screen shot shows the results of hitting the "auto-fill" button.

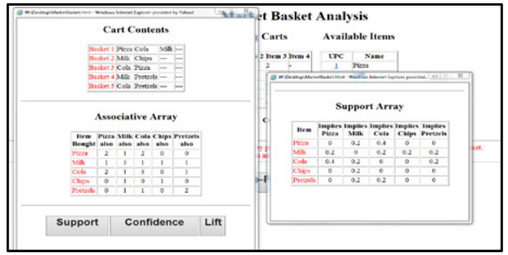


The following window opens after the student has hit the calculate button. The windows shows the shopping cart contents (in words) and the calculated associative array.

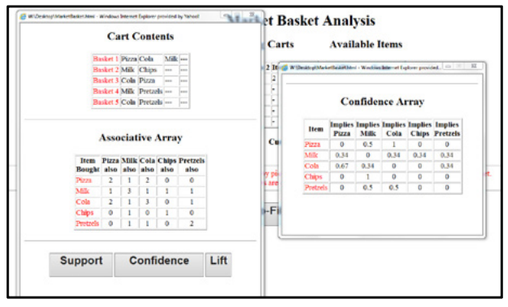


Next the students can hit buttons to calculate and display the lift, confidence, and support arrays, and this is illustrated in the following screen shots.

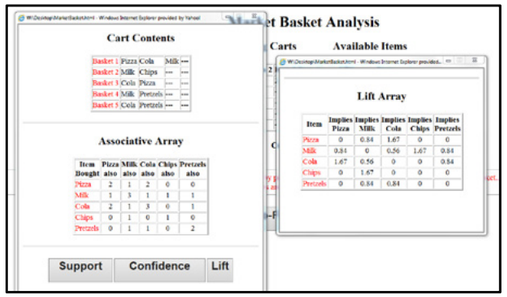
Support



Confidence



Lift



One effective way to use the tool in the classroom is to ask several students to place several items in their basket based upon what they commonly buy at a convenience store. After one has those answers from several students, then the tool is used to perform and display the results. One often get interesting and unexpected results.

CONCLUSION

This paper has described a general approach that was developed to provide intuitive and interactive learning of data analytics core principles. This has proven very useful in practice, particularly for the student's understanding of the application, and being able to engage the student with the topic.

REFERENCES

Brandon, D. "Teaching Data Analytics Across The Computing Curricula", The Journal of Computing Sciences in Colleges, Volume 30, Number 5, May 2015

Computerworld, The 2014 Premier 100 IT Leaders, February, 2014

Gartner Reports. "Business Intelligence/Analytics Is Top Area for CFO Technology Investment Through 2014", May 2013

Han, J. and Kamber, M. (2011) Data Mining: Concepts and Techniques, Morgan Kaufmann, NY, NY

Information Week, Research: Big Data and Analytics Staffing Survey, October 2012

The Economist, June 10, 2004. "A Golden Vein"

The Economist, February 27, 2012. "Data, Data Everywhere"