Knowledge-driven DSS and Data Mining: What is the "true story" about using data mining to identify a relation between sales of beer and diapers?

This is one of those recurring questions related to a famous decision support example. The story of using data mining to find a relation between "beer and diapers" is told, retold and added to like any other legend or "tall tale". I can't recall exactly when I first heard a version of the tale, but I have used the story and added to it myself on occasion. The following are some versions of the tale.

An article in The Financial Times of London (Feb. 7, 1996) stated, "The oft-quoted example of what data mining can achieve is the case of a large US supermarket chain which discovered a strong association for many customers between a brand of babies nappies (diapers) and a brand of beer. Most customers who bought the nappies also bought the beer. The best hypothesizers in the world would find it difficult to propose this combination but data mining showed it existed, and the retail outlet was able to exploit it by moving the products closer together on the shelves."

Bill Palace at UCLA (Spring 1996) in his web lecture notes writes "For example, one Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure beer and diapers were sold at full price on Thursdays."

Hermiz and Manganaris (1999) stated "One of the most repeated (though likely fabricated) data mining stories is the discovery that beer and diapers frequently appear together in a shopping basket. The explanation goes that when fathers are sent out on an errand to buy diapers, they often purchase a six-pack of their favorite beer as a reward."

Also, the 8th Annual Virginia High School Programming Contest (2001) had a problem titled Beer and Diapers. The problem statement begins "Store owners have long noticed that inspecting customer transactions can increase their profit. For example, placing the items frequently purchased together next to each other can stimulate purchasing of these items. Obviously, milk and cereal are frequently purchased together. However, some patterns are less obvious. For example, it was found that people who buy diapers also buy beer. Given a number of transactions, your job is to find a pair of items that frequently occur together."

You'll find other versions on the web and in data mining books. As a result of student questions and my own curiosity, I decided to try to find out the "truth" about this story. In July 2002, I received a media advisory about a live webcast on the past, present and future of data mining sponsored by Teradata, a division of NCR. The webcast was celebrating the 10th anniversary of a beer and diapers study and the data mining legend it started. I couldn't participate in the "live event" on July 31, 2002, but I did watch the archived webcast and the moderator, Holly Michael of Teradata, emailed me a transcript in September 2002.

Thomas Blischok, CEO of MindMeld, Inc., was one of the four panelists. Blischok managed the original study that started the beer and diapers legend. Holly Michael began the webcast by

summarizing the legend. In her version "A number cruncher was examining retail check-out data. He discovered a strange correlation, a higher than expected pairing of beer and diapers in afternoon transactions, and presumably the data indicated that young fathers were likely to pick up something for themselves as they picked up baby supplies on their way home from work. The story goes on to say that the retailer then rearranged the displays to boost sales of both products."

Holly then turned the webcast over to Thom Blischok who explained his early 1990s data mining project for Osco Drug. Thom noted that Osco Drug is one of the pioneering companies in data mining. He said "as we worked with the senior management team of their organization, we helped them create a totally new merchandising strategy. A merchandizing strategy which was focused on buying what was sold in the stores versus the traditional methodology at that time of selling what was bought by the buyers."

According to Blischok, "Their senior management team had a vision, and their vision was centered around a strategy to reinvent the store centered on consumer demand. This is where the legend began. We took over 1.2 million market baskets. A market basket is the stuff you put in the physical cart and check out at the register. And these represented transactions from about 25 stores. Our strategy on the NCR side was to discover what people bought in a given shopping experience."

And what about the legend? Blischok said "Yes, if we go back to the legend, we did discover that between 5:00 and 7:00 p.m. that consumers bought beer and diapers. This was an insight that the retailer had never seen before, and the fact that we discovered this affinity was not the real transformational event that occurred. What this showed Osco in this early pioneering effort was that it was possible to redesign the store based on consumer preferences at the center of all decisions. Their management team got it. They simply understood that they had the opportunity to change. Well, in reality they never did anything with beer and diapers relationships. But what they did do was to conservatively begin the reinvention of their merchandising processes."

Mike Grote, Director of the Teradata Data Mining lab in San Diego, followed up on Blischok's presentation with an update on data mining in 2002. Mike noted "So if we think back about that beer and diapers story that we are leveraging here today for purposes of the press conference, there are certainly some limitations associated with that as a data mining example, especially when contrasted with where the state of the art of data mining is today. I think in the context of that example, the tools that were state of the art, query generation tools, allowed Thom and his team to examine very, very large numbers of transactions and see where some particular purchases occurred together. So what does that show, and how would we contrast that with how we might approach the problem today? Well, probably what we would do with the problem today is we would use some additional tools that would not only enable us to identify where events were happening together, but they would in fact allow us to make determinations whether that one event led to a significantly increased likelihood that another event is going to occur, or whether one purchase significantly increases the chance that another purchase is going to happen."

Does everyone agree with the above account? YES and NO! John Earle in a note at www.riggs.com posted 12/21/1998 wrote "I worked for Teradata and the man attributed with starting the myth. We had done a data discovery for Osco Drugs...looking for affinities between what items were purchased on a single ticket. Then we suggested tests for moving merchandise in the store to see

how it affected affinities. ...Our 'fearless' leader, Thom Blischok, when talking with prospects and the press, didn't distinguish between the actual affinities tested and our hypotheses. Our job was to sell the value of systems. Sometimes in selling, fact blurred with folklore."

Tom Fawcett of HP Labs posted a note on the origin of the "diapers and beer" example at KDnuggets.com on Wednesday, June 14, 2000. Fawcett provides a third hand explanation of the origin of this example from Lounette Dyer via Ronny Kohavi. His posting claims Thom Blischok "dreamed up the 'diapers and beer' example. To the best of my knowledge it was never supported in any data that they analyzed."

Ronny Kohavi in an email at www.kdnuggets.com dated July 6, 2000 wrote "For my invited talk at ICML in 1998, I tracked the beer and diapers example further. Check out slide 21 in http://robotics.stanford.edu/~ronnyk/chasm.pdf. Basically, I found the person in Blischok's group who ran the queries. K. Heath ran self joins in SQL (1990), trying to find two item sets that have baby items, which are particularly profitable. She found this beer and diapers pattern in their data of 50 stores over a day period. When I talked to her, she mentioned that she didn't think the pattern was significant, but it was interesting."

So what are the facts? In 1992, Thomas Blischok, manager of a retail consulting group at Teradata, and his staff prepared an analysis of 1.2 million market baskets from about 25 Osco Drug stores. Database queries were developed to identify affinities. The analysis "did discover that between 5:00 and 7:00 p.m. that consumers bought beer and diapers". Osco managers did NOT exploit the beer and diapers relationship by moving the products closer together on the shelves. This decision support study was conducted using query tools to find an association. The true story is very bland compared to the legend.

So if someone asks you about the story of "data mining, beer and diapers" you now know the facts. The story most people tell is fiction and legend. You can continue telling the story, but remember no matter how you tell it, the story of "data mining, beer and diapers" is NOT a good example of the possibilities for decision support with current data mining technologies.

References

Brand, E. and R. Gerritsen, Association and Sequencing, February 1998, URL http://www.dbmsmag.com/9807m03.html.

Cohen, N., Data Mining: Nagging that it really adds up, 2000, URL http://www.open-mag.com/features/Vol_16/datamining/datamining.htm

Fawcett, Tom, Origin of "diapers and beer", posted at KDnuggets.com, Wednesday, June 14, 2000, URL http://www.kdnuggets.com/news/2000/n13/23i.html.

Fu, X., J. Budzik, K. J. Hammond, Mining Navigation History for Recommendation, Infolab, Northwestern University, in Proceedings of Intelligent User Interfaces 2000, ACM Press, 2000, URL http://dent.infolab.nwu.edu/infolab/downloads/papers/paper10081.pdf.

Hermiz, K. and S. Manganaris, Beyond Beer and Diapers, DB2 Magazine, Winter 1999, URL http://www.db2mag.com:8080/db_area/archives/1999/q4/miner.shtml.

Kohavi, R., Origin of "diapers and beer", email dated July 6, 2000, http://www.kdnuggets.com/news/2000/n14/8i.html.

Michael, H., Transcript of the Beer and Diapers webcast, email, September 3, 2002.

Palace, Bill, Data Mining, a technology note prepared for Management 274A, Anderson Graduate School of Management at UCLA, Spring 1996, URL http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm

Riggs Eckelberry's OF INTEREST, More On Diapers and Beer, Monday, December 21, 1998, URL http://www.riggs.com/archives/1998_12_01_OIarchive.html.

Teradata Webcast, Beyond Beer and Diapers - The Origins and Future of Data Mining, archived 7/31/2002 at Teradata.com.

The above response is from Power, D., What is the "true story" about data mining, beer and diapers? DSS News, Vol. 3, No. 23, November 10, 2002.

Author: Daniel Power
Last update: 2005-08-06 21:53