ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
December 2014,
Vol. 7, No. (3):
Pgs. 316-322

Building Data Mining For Phone Business

AKAZUE MAUREEN and OJEME BLESSING

Mathematics & Computer Science Department, Delta State University, Abraka, Nigeria.

(Received: November 16, 2014; Accepted: December 10, 2014)

ABSTRACT

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. A framework to guide a phone business is discussed using data mining tools (decision Tree) to predict future trends and behaviors of their customers, thus, allowing their businesses to make proactive, knowledge-driven decisions. The impact of integrating data mining with acquisition marketing campaign management is also explained.

Key words: Data Warehouses, Database Marketing, Data Mining, Attrition (churn), Campaign Management, Customer segment.

INTRODUCTION

We are in an age often referred to as the information age because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos led to the creation of structured databases and database management systems (DBMS). The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed (Osmar, 1999).

The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information. Today, we have far more information than we can handle: from business transactions and scientific data, Medical and personal data, Surveillance video and pictures, to satellite pictures, Satellite sensing, Games, Digital media , Virtual Worlds , CAD and Software engineering data, text reports, the World Wide Web repositories and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data.

Therefore, data mining techniques are the result of a long process of research and product development (Elder et al, 1998). This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments.

Data mining has emerged as one of the key features of many homeland security initiatives as a means for detecting fraud, assessing risk, and product retailing. In the context of homeland security, data mining is often viewed as a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records (Thearling, 1998).

While data mining represents a significant advance in the type of analytical tools currently available, there are limitations to its capability. One limitation is that, it does not tell the user the value or significance of these patterns. Hence, these types of determinations must be made by the user. A second limitation is that, while data mining can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship (Seifert, 2004).

MATERIALS AND METHOD

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include, sequence or path, classification, clustering, and forecasting.

As an application, compared to other data analysis applications such as structured queries (used in many commercial databases) or statistical analysis software, data mining represents a difference of kind rather than degree. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis. The two precursors are important for a successful data mining exercise; a clear formulation of the problem to be solved, and access to the relevant data.

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased ability to combine data from disparate sources into a single searchable source. Organizations use data mining as a tool to survey customer information, reduce fraud and waste, and assist in medical research. However, the proliferation of data mining has raised some implementation and oversight issues as well. These include concerns about the quality of the data being analyzed, the interoperability of the databases and software between agencies, and potential infringements on privacy.

Motivation

Data mining software allows users to analyze large databases to solve business decision problems. In some ways, data mining is an extension of statistics, with a few artificial intelligence and machine learning twists thrown in. Like statistics, data mining is not a business solution, it is just a technology. Data mining can do the job of segmenting prospective customers (Greening, 2006) as well as Customer relationship management (CRM), which is a process that manages the interactions between a company and its customers (Thearling, 2007), maximizing the Value of Interacting with Your Customers (Thearling, 2001), and customer acquisition (Berson et al, 1998).

The Two Crows Corporation (1996) stated that, Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

Data mining extracts information from a database that the user did not know existed. Relationships between variables and customer behaviors that are non-intuitive are the jewels that data mining hopes to find. And because the user does not know beforehand what the data mining process has discovered, it is a much bigger leap to take the output of the system and translate it into a solution to a business problem.

Therefore, the objectives of this research is to elicit the value of data mining and recommend follow-up data mining initiatives, to drawn out how data mining applications automate the process of searching the mountains of data to find patterns that are good predictors of purchasing behaviors and finally, to design a framework that will guide a mobile phone company to know which customers

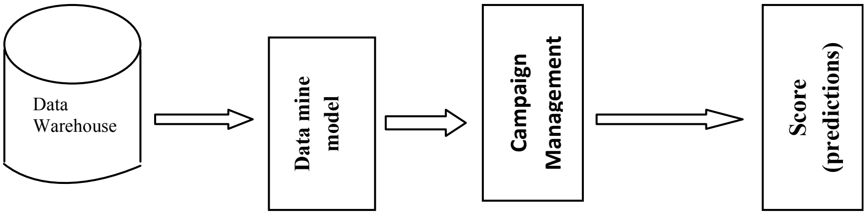

Fig. 1: Role of Campaign Management software

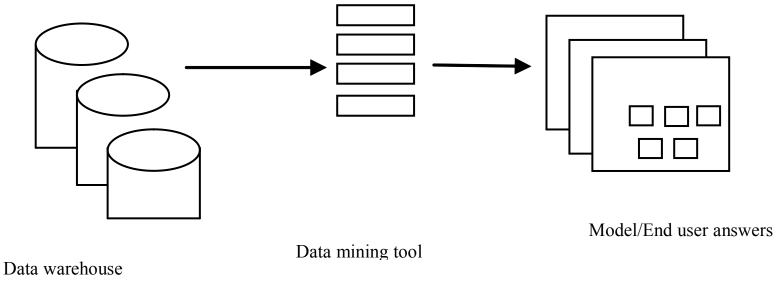

Fig. 2: The Data Mining Model

might be interested in a promotion and the eligibility of that customer through the use of decision tree.

Customer Acquisition Concepts

Before the process of customer acquisition begins, it is important to think about the goals of the marketing campaign. In most situations, the goal of an acquisition marketing campaign is to turn a group of potential customers into actual customers of your product or service.

- It is a good start when customers ask for more information about your products or services. It could signal the beginning of a long-term customer relationship. You might also want to track conversions, which are follow-ups to inquiries that result in the purchase of a product. There are two kinds of negative responses: rejections and non-responses. Rejections, by their nature, correspond to specific records in the database that indicate the negative customer response. Non-responses, on the other hand, typically do not represent records in the database. Non-responses usually correspond to the absence of a response behavior record in the database for customers who received the offer.
- Most acquisition marketing campaigns begin with the prospect list and sometimes, it is necessary to add additional information to a prospect list by overlaying data from other sources. More complicated overlays are also possible such as matching customer against purchase, response, and other detailed data that the data vendors collect and refine.
- Once you have a list of prospect customers, you will need to send out a test campaign in order to collect data for analysis. Besides the customers you have selected for your prospect list, it is important to include some other customers in the campaign, so that the data is as rich as possible for future analysis.
- This random selection should constitute only a small percentage of the overall marketing campaign, but it will provide valuable information for data mining.
- More sophisticated techniques than random selection do exist, such as those found in statistical experiment design and multi-variable testing (MVT). Although this circular process (customer interaction? data collection? data mining? customer interaction) exists in almost every application of data mining to marketing, there is more room for refinement in customer acquisition campaigns.
- Once you have started your test campaign, the job of collecting and categorizing the response behaviors begins. Immediately after the campaign offers go out, you need to track responses. So, with the test campaign response data in hand, the actual mining of customer response behaviors can begin. The target variable that the data mining

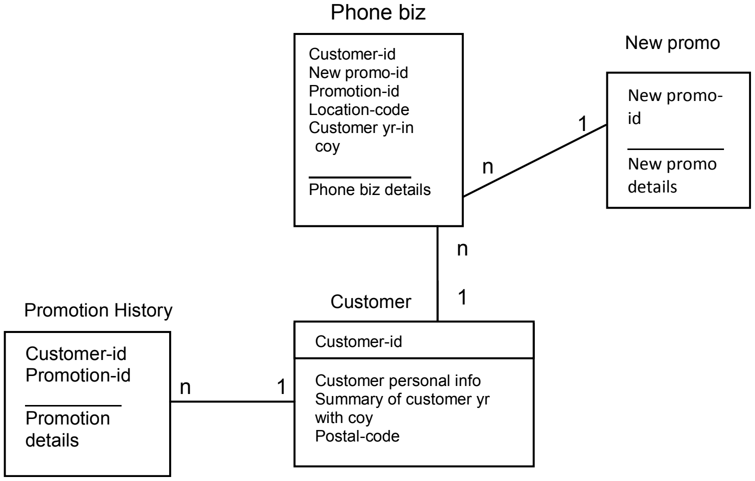

Fig. 3: Shows the data model used as the primary source of data for this study

software will predict is the response behavior type at the level you have chosen (binary or categorical).

In the end, a model (or models, if you are predicting multiple categorical response behaviors) will be produced that will predict the response behaviors that you are interested in.

Research methodology

An Overview of Data Mining Techniques

The process of data mining consists of three stages: (1) the initial exploration, (2) model building or pattern identification with validation/verification, and (3) deployment. The exploration stage usually starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records. The Model building and validation stage involves considering various models and choosing the best one based on their predictive performance (i.e., explaining the variability in question and producing stable results across samples. The deployment (final) stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

Among the most common data mining algorithms in use today (Thearling, 2007), we chose to use a decision tree as a predictive model because from a business perspective, decision trees can be viewed as creating a segmentation of the original dataset (each segment would be one of the leaves of the tree). Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. The other reason why decision tree is chosen is because it's a favored technique for building understandable models. Because of this clarity, they also allow for more complex profit and ROI models to be added easily in on top of the predictive model as well as their high level of automation and the ease of translating, decision tree models into SQL for deployment in relational databases.

CHAID or Chi-Square Automatic Interaction Detector and CART (Classification and Regression Trees) are popularly used predictive algorithm. CART picks the questions in a much unsophisticated way: It tries them all. After it has tried them all CART picks the best one, uses it to split the data into two more organized segments and then again asks all possible questions on each of those new segments individually.

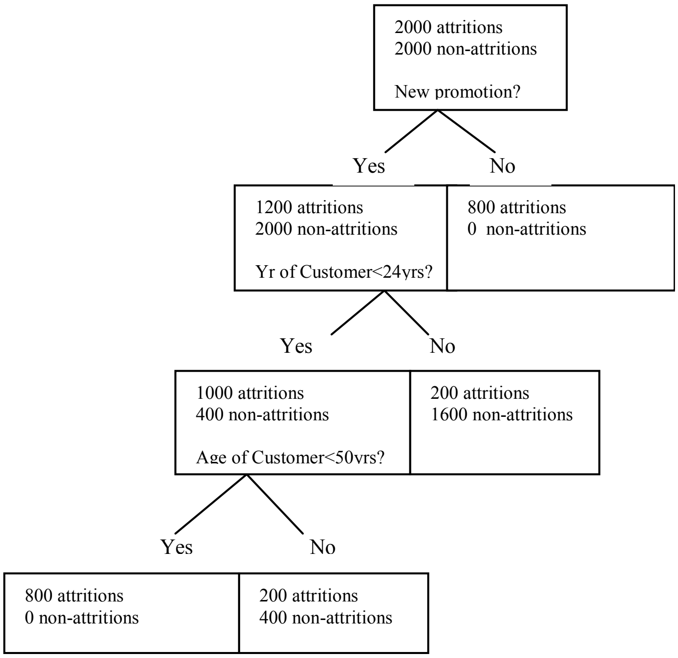

Fig. 4: A decision tree model that makes a prediction on the basis of a series of decision

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting.

A Data Mining Framework that will guide a phone business

The Promotion Program categorized the customers into categories such as "yr of customer", age, promotion type, gender, and so on. This is used to join the promotion data to the customer file to create the clustering input record required by the decision tree used. This code illustrates how the transaction data is pivoted, aggregated, and inserted into each customer record.

Data cleansing

In the phone business study, all the promotion variables (amount recharge, and number of promotions undertaken by a customer) were assigned zero when null using the following SQL code:

```
update target set cat1_phone_q2=0
where cat1_phone_q2 is null;
update target set cat1_phone_q2=0
where cat1_promotion_q2 is null;
update target set cat1_amt recharge_q2=0
where cat1_amt recharge_q2 is null;
```

All categorical variables were made consistent, and unknown values were assigned a code with SQL similar to the following:

```
update target set gender='U'
where gender is null or gender=' ';
```

When coding missing values, a binary field for all missing categorical fields and numeric fields were created using SQL code similar to the following:

```
alter table
target add unknowngender smallint;
update target set unknowngender=1
where gender='U';
update target set unknowngender=0
where unknowngender is null;
```

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. For instance if we were going to classify customers who churn (don't renew their phone contracts or participates in promotions) in the Phone Business, a decision tree might look something like that found in figure 4. It is such that:

- It divides up the data on each branch point without losing any of the data (the number of total records in a given parent node is equal to the sum of the records contained in its two children).
- The number of attritions and non-attritions is conserved as you move up or down the tree
- It is pretty easy to understand how the model is being built (in contrast to the models from neural networks or from standard statistics).
- It would also be pretty easy to use this model if you actually had to target those customers that are likely to churn with a targeted marketing offer.

From here, some intuitions about the customer base were built. E.g. "customers who have been with the phone business for a couple of years and have up to date participated in promotions are pretty loyal". The output of the model, the prediction, is called a score which is typically a numerical value that is assigned to each record in the database and indicates the likelihood that the customer whose record has been scored will exhibit a particular behaviour.

CONCLUSION

The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Business questions that traditionally were too time consuming are answered by Data mining tools.

As long as the necessary information exists in a database, the Data Mining process can model virtually any customer activity. The key is to find patterns relevant to current business problems. The typical questions that Data Mining answered includes: customer that is most likely to participate in the mobile-phone promotion service, the probability that a customer will purchase at least N15000 worth of recharge card, and the age grades that are most likely to respond to a particular offer or promotion? Answers to these questions helped to retaining customers and increased campaign response rates, which, in turn, increase buying, cross-selling and return on investment (ROI).

REFERENCES

1. Berson, A., Smith, S., and Thearling, K. An Overview of Data Mining Techniques http://www.amason.com/exec/obidos/ISBN+0071344446/kurthearlingdatA/ obtained from web- 2007 (1998).
2. Elder, J.F., Abbott, D.W. (1998). Fourth International Conference on Knowledge Discovery & Data Mining. http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics_bw.pdf.
3. Greening, D.R. (2006). Data Mining on the Web: There's Gold in that Mountain of Data. http://www.newarchitectmag.com/documents/s=5339/new1013637424/mailto:greening@andromedia.com
4. Osmar, R. Z. (1999). Principles of Knowledge Discovery in Database Introduction to Data Mining http://www.exinfm.com/pdffiles/intro_dm.pdf.
5. Seifert, J.W., and Relyea, H.C. (2004). Information Sharing for Homeland Security: A Brief Overview. CRS Report RL32597.). http://www.fas.org/irp/crs/RL31798.pdf
6. Thearling, K. (1998). An Overview of Data Mining at Dun & Bradstreet www.thearling.com/text/dsstar/privacy.htm
7. Thearling, K. (2001). Campaign Optimization: Maximizing the Value of Interacting with Your Customers. www.thearling.com/text/optimazation/optimazation.htm
8. Thearling, K. (2007). An Introduction to Data Mining : Discovering hidden value in your data warehouse. http://www.thearling.com/text/dmwhite/dmwhite.htm
9. Two Crows Corporation. (1996). Introduction to Data Mining and Knowledge Discovery, Third Edition. New York: Addison Wesley.