**MGS 8040 Data Mining Overview**

**Dr. Satish Nargundkar**

**Introduction:** Data mining is the process of selecting, analyzing, and evaluating data collected from different sources within an entity’s database. The main objective of conducing data mining is to help develop and support business strategies that can enhance the performance of an entity through increasing revenue, decreasing costs, or a combination of both. Data mining is now used across a wide range of industries including retail, finance, manufacturing, healthcare, and hospitality. Data mining techniques go beyond simply reporting into some form of predictive analysis, allowing us to see certain facts, relationships, trends, and patterns that can be used to take action.

The key data mining tasks are Prediction/Classification, Segmentation, and Association.

**Prediction/Classification:**  Based on historical data, we can make certain predictions of future behaviors. Customers (or potential customers) may be classified into groups based on a prediction of their likelihood of responding to product offers. There are several other possible customer behavior models, such as the likelihood of default (typical credit scoring models), likelihood of churn, or the likelihood of fraud. Healthcare applications may include classification of patients into risk categories for certain diseases. Law enforcement can use data to look for patterns of behavior to identify potential threats.

The mathematical techniques used for prediction/classification include Multiple Regression, Discriminant Analysis, Logistic Regression, and Artificial Neural Networks.

**Segmentation:**  Segmentation is the process of dividing a large group of entities into smaller, more homogenous groups, based on characteristics deemed relevant by the analyst or decision maker. Market segmentation is the most common business application, so that a business can target products to groups most likely to be interested in purchasing them. Segmentation techniques may be used in the pharmaceutical industry to group different types of chemical molecules into clusters, or by astronomers to group together different kinds of celestial objects based on their properties. In a different type of grouping, various measures of an underlying concept, such as question items on a survey that all measure “Customer Satisfaction” in various ways, may also be grouped into segments.

The mathematical techniques used for segmentation are Cluster Analysis and Factor Analysis, although segmentation can be done subjectively in some cases. Also, with categorical data, such as gender, the categories are automatically segments.

**Association:** Amazon Inc. made this technique popular with their recommendations to customers for products they might like to purchase, based on their current purchase or interest. Netflix also uses the same approach - people who want to watch *Captain America* may also want to watch *Iron Man*, and therefore, the website will suggest *Iron Man* as a cross-selling item during the customer interaction. E-Harmony, Match.com, and other social sites use the same approach of finding associations among people based on their characteristics.

The techniques used for association are generally matching or memory based techniques, and what is known as Market Basket Analysis.

**An Application in the Financial Services Industry**

Banks and other financial service companies have long used some of the techniques mentioned above. Consider the typical customer life cycle for a financial product, such as a credit card or a loan, shown in figure 1 below.

**Figure 1:** Customer life cycle, financial services

As the figure shows, the bank would go through the following four stages:

**Product Planning:** The bank determines product characteristics in order to initiate a target marketing effort to match the new product with appropriate customers. It is important to ask the following questions: Who will the bank be offering the credit card to (income level, demographics, and educational level)? Should the bank charge an annual fee? What kinds of features will the credit card carry?

**Customer Acquisition:** Take actions to acquire potential customers in a cost-effective way. The bank can use statistical models to predict response rate; hence, the bank can budget its marketing expenses accordingly. Risk models (credit reports/scores) can help the bank determine the riskiness of the credit card applicants, allowing them to set different interest rates or credit limits, thereby controlling the amount of bad debts due to write-offs.

**Customer Management:** Once a customer is acquired, retaining the right ones requires further analysis of data. Monthly transaction and billing data provides a rich source of information on customer behavior. Behavior models to predict customers’ future behaviors are commonly used. For example, the bank can predict the likelihood of further delinquency of customers that miss a payment. The level of risk determines the kind of action that can be taken, from doing nothing all the way to shutting down an account. The bank can also use behavior models to figure out ways to retain valuable customers, and prevent churn by providing the right incentives.

**Collections and Recovery:** If bad debts arise, the bank will have to figure ways to recover those bad debts. For example, the bank could sell the ownership of those receivables to a third party by offering a price discount. These bad debts provide information to further improve the risk models in the next cycle.

**The Data Mining Process**

The following are the key steps that you, as an analyst, would follow in the data mining process:

Determine Business Requirements

Understand, Collect, Aggregate, and Clean the Data

Develop Appropriate Models

Evaluate the Models - Validation

Monitor/ Control

Implement

**Figure 2:** The Data Mining Process.

**Understanding Business Needs:** Developing models to help a business requires first an understanding of the business itself. What is the goal of the model? What kind of decision making will it support?

**Data Understanding:**  Modern statistical software is powerful enough to make model building easy, and can lull the analyst into believing that all it takes is for data to be thrown into the computer, and it will do the rest! Without a good understanding of data, we are likely to get meaningless models. Business and data understanding go hand in hand. When we understand the domain, we are likely to know whether or not the data make sense. There are various reasons for data to be in error, and recognizing bad data when you see it is critical. Also, the analyst must know the level of aggregation necessary. For instance, individual transaction patterns may be needed to detect fraudulent transactions, while monthly aggregates might the necessary for developing a model to predict customer default risk.

Data must be modified in many cases in order to be usable. Cleaning, aggregation, and standardization are some of the tasks that may be performed before modeling. For example, since companies’ sizes are not equal, financial ratios are widely used by analysts in order to objectively assess companies’ performances. The raw data may have the component variables. The analyst must compute the ratios needed for analysis.

**Modeling, Evaluating, and Deployment:**  Once we have all the useful data, we can start developing models, a process that is relatively simple due to sophisticated software. Evaluation of models is done typically by using a validation (holdout) sample of data and checking if the model performs well on it. Implementation of the model is followed by monitoring its performance in a real business scenario, and updating the model as needed.

**Discussion:** Can we always rely on models? What happened with sub-prime loans in the mortgage crisis? Did their “risk models” work?

**Course Project:**

You are responsible for working on a course project in a team. Your team will analyze data from any industry, build a model using a categorical dependent variable to predict something and classify entities into groups based on those predictions. You will demonstrate that the model is valid, and include sample monitoring reports. You can discuss alternate project ideas with me.