Pharmaceutical Market Share Analysis using IBM s Data Mining Capabilities

Discovering Dominant Patterns that Predict Market Share.

Ashok N. Srivastava, Ph.D., Principal Investigator

Dion Cummings, Ph.D., Data Mining Analyst

IBM Global Business Intelligence Solutions

Note from the Authors

This document contains excerpts from the final report that we delivered to a large pharmaceutical company. Confidential results have been omitted, as has specific information about the drugs and drug types, and the company name. Please direct any questions to the first author.

Abstract

The drug market is a very competitive market place because of the large number of drugs being introduced to the market place by drug companies vying for market share. One of the key components for success in claiming or preserving market share is the performance of its sales force.

A large pharmaceutical Company hired IBM Global Business Intelligence Solutions (IBM GBIS) Data Mining Analysts to apply data mining technology to quickly and accurately establish a relationship between sales force activities and market share of their drugs. This involved identifying dominant patterns that predict market share of the drugs.

Tools from IBM s Business Discovery Solution (BDS) are applied to address this problem. This document discusses the methodology used to identify key variables that appropriately predict market share. We find that with appropriate data transformations, we are able to build models that account for a reasonable proportion of the variability in the data that can predict market share.

Introduction

Pharmaceutical companies manufacture several kinds of drugs that fall under several product types. These drug types makeup the company s total drugs market share. Currently, many pharmaceuticals only track sales force activities associated with the selling of two drug types since these two drug types account for most of their market share.

Like many drug companies, one of the ways the company introduces its products to the medical community is through its sales force. Calls are made to a network of doctors nationwide, in an effort to demonstrate its products. Another method used for introducing these products is by hosting symposiums and dinners where prescribers are educated about these drugs and the benefits it holds for their patients.

Regardless of the methods used for product introduction, this pharmaceutical company, like many other drug companies gives many complimentary samples annually, with the hope of promoting sales.

The key to how well a product sells depends largely on the sales force in terms of how well they perform their sales activities. A sales team may make a large number of calls, perform many shows and leave many samples of a product, but, that does not guarantee that a prescriber will recommend that product to its patients in every instance. Market share size for Pharmaceutical drugs is based in part on the performance of the sales force, and how well its products are received by the medical community.

The purpose of this document is two-fold: first, to deliver the results of the data mining activity. Second, to outline the data mining methodology, so that the ideas may be applied to future market share prediction problems at a pharmaceutical company.

Goals of a Predictive Data Mining Model

In general, a predictive data-mining model can have two possible uses. The first use falls under the rubric of *analysis*, where the model is used to obtain key predictors of the target variable. In the current study, the target variable is market share of Pharmaceutical drugs, and the analysis of the model leads to key predictors of market share. Given this information, a marketing analyst can use these key predictors to further verify the plausibility of the predictors and to take action to optimize the sales force activities.

The second general use of such a model falls under the rubric of *prediction*, where the goal is to obtain high accuracy predictions. Traditionally, this use follows once the model has been heavily tested on a number of different data sets with different characteristics. In the current study, the predictive models, once appropriately validated, could be applied to production data.

Domain Specific Data Mining Issues

Perhaps the single most important issue that arises in this study is that the number of data points since we are only dealing with 24 months of data. This adds a high degree of complexity to the prediction task because the number of data points is very small. Since any predictive algorithm relies on examples from historical databases to predict future behavior, a small number of data points can make generalization for future behavior difficult. To mitigate this issue, we implemented a data processing strategy to increase the number of data points for the purpose of the analysis.

Pharmaceutical Database

The success of the data-mining project depended upon fast and easy access to a database that contains information relevant to the project. The pharmaceutical gave us access to files from their databases which was made available on one of IBM s SP2 machines in Dallas (dmine4). These files ranged in sizes from approximately 1MB — 1.5GB and consist of 24 months of historical data.

Our results included predictions of the market share based on 24 months of aggregated data by product type, i.e., each month was an observation which was the sum of all related physician s data, by product type.

Raw Data Distributional Analysis

Preliminary analysis was performed on the data sets using SAS to obtain an understanding of the distributions of the variables. Although data mining algorithms tend to be robust to the distribution of the variables, these analyses gave us key insights into the frequency of missing or invalid values and the need to replace them. Because of the nature of the study, replacement of missing and/or invalid data was addressed at the aggregate level.

Data Preprocessing

Most of the time spent on this project was devoted to the preprocessing of the data necessary for model building. This was not a trivial matter because of the data layout and the file sizes. One of the main obstacles that we had to overcome in processing these files was out of memory conditions, since most UNIX file systems have a 2 GB memory limit. This condition is very common during sort procedures when large sizes of data are used. For example, sorting a file that is 1/2 GB can require more than 4 GB of memory to finish the process.

To obtain the data necessary for this study, a series of joins, aggregations and extraction had to be performed. Note that before aggregation took place, duplication had to be eliminated.

The data necessary for this study was split over 10 files. To access it, these files had to be merged - in some cases by physician reference number or by plan code or by prescriber location.

Data Quality: Missing Values

Missing and outlier values pose an important problem in most data mining projects. In this project, there were several months with missing data and several outlier points. Because of the limited number of data points, the missing and outlier values were dealt with using linear regression.

Section 2: Market Share Definition and Calculations

Market share was computed as the ratio of the number of scripts of a particular product to the total number of products sold. The following gives the definition of market share that we used. This definition was checked and validated by Pharmaceutical domain experts.

Definitions:

Let:
d = index of doctors (d = 1...total number of doctors)
p = index of products (p = 1 ... total number of products)
P = total number of products
t = index of time periods (t = 1...total number of time periods)
T = total number of time periods
N(t) = number of products sold by all doctors.
L(p,t) = number of Pharmaceutical products of type p sold in time period t.
share(p, t) = market share for the pth product at time period t.

Market share of product p can be calculated as follows:

share(p, t) = L(p, t)/ N(t)

Data Transformation

In order to make reliable predictions additional variables were needed. This meant creating derived variables for the predictions. Because we were dealing with time series data, an auto regressive model type data was appropriate.

Auto regressive models are models where the independent variables are all previous values of the same time series. For example, if the time series values are denoted by Y_1 , Y_2 , Y_n , then with a dependent variable Y_t we might try to find an estimated regression equation relating Y_t to the most recent time series values Y_{t-1} , Y_{t-2} , and so on. These independent variables are also called lag variables.

Section 3: Results

Section 3 presents the results of the studies performed. The first topic for discussion is the segmentation results, followed by the prediction results. Data replication strategy is used to generate more data for analysis.

Exploratory Model: Segmentation Results

The purpose of the segmentation algorithm is to identify subsets of the records in the data set that have similar characteristics. Segmentation is an important method for obtaining both a global understanding as well as a local understanding of the database. *Global understanding*, in this case, refers to an understanding of the trends of the entire database, whereas, *local understanding* refers to an understanding of the characteristics of specific subsets of the database.

Business Motivation for Segmentation

Although the segmentation data-mining algorithm can have many general applications, there are several specific reasons why the segmentation algorithm is applied in this study, a few of which are listed here:

1. Segmentation gives us the ability to understand which market share parameter values tend to occur together and to what degree they co-occur. We look to segmentation to give us other, less obvious understandings.

2. Segmentation can give key insights into the characteristics of high and low market share.

3. Segmentation can quickly identify small subsets of parameters for high and low market share that have similar characteristics and would fall through an ordinary analysis.

Predictive Models: Radial Basis Function (RBF) Results

The purpose of a predictive model is to predict a target quantity (in this case market share) given other information (sometimes called input information). The other information may or may not have a predictive value with respect to the target variable. Part of the task of a predictive model is to identify what are the relevant predictors of the target given the target variable. The other task is to produce high-quality predictions, meaning that the predictions should be made with high accuracy. The RBF is the algorithm that we used in this study because all of the input variables are continuous. The following subsection outlines the business motivation for using this algorithm.

Business Motivation for RBF

Although the BDS software package contains several predictive algorithms (radial basis functions, neural networks, and decision trees) after understanding the requirements of the domain experts, it became clear that the RBF algorithm would be most suited for predicting market share. RBF segments your data so that cluster membership allows accurate prediction of the target field. It allows modeling of multiple modes and produces a segment and descriptions of the segment.

The decision tree is less accurate and not as suited for predicting continuous variables as RBFs. Neural networks are usually more accurate that RBF but are considerably slower. In our study, one neural net model was constructed, but it did not offer any advantage over RBF.

Section 4: Discussion and Conclusions

The main results of our study found a negative correlation among several market share variables that can possibly be attributed to some type of sampling effect where doctors that are under selling are targeted. In other words, the marketers are selling more to the doctors that are under selling Pharmaceutical products.

The main result of this study demonstrates a procedure to identify likely predictors of market share. These predictors are derived quickly and directly from the data. The predictive power of the models might be higher if more time related variables were available. The models developed consistently identify key predictors that can be verified by domain experts as being viable market share predictors. The methodology provided in this study demonstrate a way for the sales force activities to be modeled and understood.