SD-10

Where Should I Dig? What to do Before Mining Your Data

Stephanie R. Thompson, Datamum, Memphis, TN

ABSTRACT

Data mining involves large amounts of data from many sources. In order to successfully extract knowledge from data, you need to do a bit of work before running models. This paper covers selecting your target and data preparation. You want to make sure you find golden nuggets and not pyrite. The work done up front will make sure your panning yields results and is not just a trip down an empty shaft.

INTRODUCTION

Data mining has garnered a lot of attention lately and for good reason. Pulling the information out of your data can give you a competitive edge, generate more profit, help you retain customers, and help you evaluate the effectiveness of advertising campaigns. As the name implies, a lot of data are involved and you need to sift through the results to find what is worthwhile. However, you cannot simply put all of your data into a SAS® dataset, click run, and have all of the answers waiting for you after lunch. Unfortunately, it is just not that easy. With a little up-front planning, you can successfully answer your business questions and generate results you can depend on.

Determining what you want to model is a good first step. Next, you need to see what data are available to contribute to the effort. Once you find the data, it needs to be packed up and put together in a way that makes sense. Lastly, you will want to do a few checks on your data quality and think about how you plan to run the models. These preliminary steps will ensure you are setting yourself up to get the best possible results from your modeling effort. Keep in mind though that not every mine strikes gold. By getting the steps right on the front end, you will be in a position to say with confidence that more data are needed, a dependable relationship between the target and data is not in the data, or how “the gold in them thar hills” just is not there. All three of these scenarios is better than handing your boss a sack of pyrite then having her take it to the bank to be told it is worthless – or worse yet that they have lost money banking on it before having it tested.

CHECKING THE MAP

The map in this case is the question you want to answer. Asking the right question can make the difference between success and failure. It is important to have a measurable result that can be quantified (e.g., how many more procedures of a certain kind are performed after training on the procedure than before training?) or categorized (e.g., are students more likely to stay enrolled vs. drop out if they receive extra math tutoring?). In both of the examples, the outcome is measurable and can be captured by data. Seek questions that you can quantify.

If your questions are too vague, you may not be able to measure them. Some questions hard to measure might be, do physicians like our company more after attending training? (How do you quantify like?), did the advertising generate more sales? (what constitutes more? One, five, 100? More compared to when – last year, last month, last week?), and does the average speed employees drive to work determine their productivity? (Can you measure this? What is the question based on? Is it positively or negatively correlated?). Without a clear definition of the outcome, you cannot generate a target to model. Even though the question sounds valid, make sure there are data to support it.

Choosing a target related to your data might generate a strong model but most likely will not tell you anything you do not already know. Examples here could be, Do student's GPA increase when they receive more A grades than other grades? (Letter grade translates to a numeric GPA), is selling price related to cost? and other similar questions. This type of question will not help you uncover relationships in the data that can help your business grow. They only reinforce something already known.

ASSEMBLING THE TOOLS

Pulling together the data can be your biggest task. Data exist everywhere in an organization, in multiple formats, and on different platforms. There may be a need to combine database data, PC files, and flat files. Are data on servers or local PCs? How do you get access to the data? Very often, you need to have approval to access data as well as proper credentials to get to it. Knowing whom to ask in a large organization can be a challenge of its own. If you cannot access the data, you cannot use it your model.

Thinking about the data you know about can seem like a large amount but it grows when you include the data you do not know about. Most companies keep their data in a central warehouse or data mart for ease of access and to ensure security. However, there are cases where things are tracked within a department or even by an individual if it is particularly sensitive information. Knowing about these sources involves asking around and looking at reports generated by the company. If you see a news release about how a new product has increased sales or that four out of 5 dentists prefer your company's product, there is data backing that up somewhere (well, hopefully, but that's another story).

If you identify data that you do not already have access to, it is a good idea to meet with the data steward/owner to explain what you would like to do. Depending on your organization, data may be viewed as a company-wide resource or still protected in individual siloes. Resistance can come from fear of misinterpretation. Allay that fear by asking questions about each data element to insure you know what it means. Headings can be misleading. Take sales for example; is it in units, retail dollars, profit margin, or some strange mix of all three? Knowing what the data mean will make for a better model and build a relationship with the data owner.

With data mining, more data are better. If you knew the relationship between adoption of a procedure and training, you would not be doing a data mining project. Get everything you can related to your target and can be linked together. The relationship may only exist for certain sub groups of the population. You need a way for the model to create the segments.

LOADING THE MULE

Knowing how to put the data together is the next step. Everything needs to be compiled and related. Using student data, for example, you may need a student ID number to join admissions, registrar, and financial aid data together. If you do not have a unique ID, do you have another way to match people up? Can you use first name, last name and birthdate or is more information needed? If you cannot bring the data together, you will not be able to model it. Here again it is important to understand your data sources. Understanding whether employee ID is the same as person ID and different from customer ID can make a huge difference in your results. Can you imagine modeling patient data that has been linked to supplier accounts payable information if you assumed the eight digit IDs represented the same thing?

Once you know how to put the data together, you need to see where it needs to be flattened. One record per student is needed for your model. If the data for race and ethnicity have multiple rows per student, it needs to be flattened before combining it with other information. Here is it also good to make sure that you truly have individuals. Is Dr. Robert Smith in Baton Rouge the same as Dr. Bob Smith in Baton Rouge?

Another thing to check at this point is whether the variables are in the correct data type. Age should be numeric if you are looking at its impact as a continuous variable. If age is stored as a character value, you will not see this effect in the model. A five-digit zip code, on the other hand, even though it is numeric represents a classification as opposed to an ordinal relationship. Making sure each is used correctly in the model is important. Spending the time up-front to get the data right is worth the time. As the saying goes, "Junk in, junk out." Getting various data sources assembled and merged for a large data mining project will take time.

SETTING UP THE SLUCE

Data quality is sometimes overlooked in the modeling process. After spending all of the time to assemble and merge the data you are ready to get on with the mining. Spending a bit more time can save you headaches later. First you want to see how complete you data are. Are there variables where the majority of the values are missing? If so, is this a problem? Do you need to impute values? Should you eliminate the variable from the model? There are no pre-defined answers to these questions as each situation is unique. Here the motto, "Know thy data" comes in to play. If you are not sure, do some digging and find out.

Another check is on the distribution of values in a variable. If you have a variable representing age, what is the average and extreme values? If you see negative values or extremely high values there is a problem that is easily spotted. A little more difficult is looking at the average. If, after fixing any errors in the data, you see an average of 46 you need to determine if that is appropriate. If the data represents a population of dentists, you might be in good shape. When modeling pediatric data, this average age would indicate a problem.

Categorical variables also need to be checked. If the choices on a survey are A, B, C, or D, you should not see any X, W, or R values in the data. If you do, you need to determine how they got into the data and if the data are valid. Case is another attribute of the data to check. In coding gender, "F" and "f" may both represent female. If they are not combined, the model will treat them as two distinct levels. As with numeric variables, knowing what to expect in categorical variables is important. If you do not know what to expect you will never know if you have bad data.

One last thing to check in the data is the number of levels of a variable. The individual ID should be unique and is not modeled. However, if you have other values in the model unique only to one row or individual, more than likely it will

not contribute to the model. Maybe you have zip code in your data. If the data are for the entire United States, that may be too granular. Could you roll up to the state level or is a regional definition more appropriate? If modeling an individual county, zip code may just be what you need.

KNOWING WHEN TO PACK IT UP

There are times that data mining will not yield a result that you can use. This does not mean that the model or data are necessarily bad. Sometimes an underlying relationship is just not there. It is better to admit this than to force a result that is not right. Maybe more data are needed, more measurements need to be taken, or the question cannot be answered as formulated. It is not failure, but rather a success you were not expecting. Honestly presenting these results will generate confidence in the work and the tool. As long as you prepared correctly up front, a model that does not come together is just that.

The answer may be down the road or it might be better to reframe the question into something else that may result in a model. Realizing that the answer is not in what you have helps you to move on to something that will be productive.

CONCLUSION

Doing some up-front work will help ensure you have a successful data mining project. By its very nature, data mining involves large amounts of information that need to be combined, checked for completeness, and evaluated for quality. Depending on the project, this work can take longer than the modeling itself but will ensure a result that you can have confidence in. Learning to formulate the question in a quantifiable manner will get you further down the road to the mine. It is important to have a question that can be addressed with the data you have. If not, you may need to hold off on modeling until you can get the right data.

Lastly, remember that not all models will yield an actionable result. There may, in fact, not be a relationship between the target and your data. That does not represent failure. It tells you something that you did not know about your business. Maybe the long held belief in the location of the lost mine is just fiction. The real gold may be hidden somewhere else. Now your task can be to find SESUGwhere it truly is.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Stephanie R. Thompson
Enterprise: Datamum
Work Phone: 901-326-0030
E-mail: stephanie@datamum.com
Website: www.datamum.com
Linked In: Stephanie Thompson, Analytics Professional
Twitter: @SRT_SESUG
sasCommunity User Name: Stephanie