Paper 127-2010

More than Models: The Data Mining Team

Stephanie R. Thompson, Rochester Institute of Technology, Rochester, NY

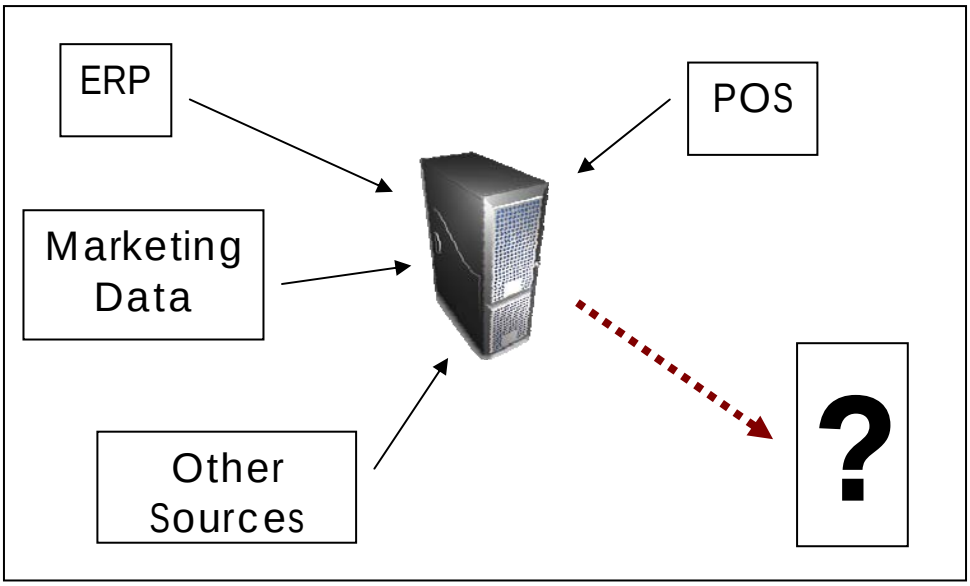
ABSTRACT

Gathering data from various sources, preparing it for modeling, imputing, partitioning, testing various models, choosing the best, presenting it to your boss, then failure? Data mining is about more than just variables and models. Developing an understanding of each variable may take more than just running some summary statistics and deciding if it is ordinal or nominal. The need for comprehension is even more critical when the data you use are from different areas of the organization. Putting together a group of subject matter experts in the early phases of a data mining project can make a big difference in the outcome of your project. They can help you eliminate extraneous or duplicative variables and put others in context to help you better understand and interpret results. This paper will discuss how subject matter experts can aid in data mining using examples from several actual projects. Learn how to leverage the knowledge to derive a better conclusion and avoid costly errors.

INTRODUCTION

A number of years ago there was a quarterly investment analysts conference call that I listened in on. I remember how the company had deployed a new merchandising strategy based on data mining. This new strategy was going to revolutionize assortments and improve sales. The results this particular quarter were not as expected. After the company's presentation, one of the analysts asked if the new merchandising strategy had anything to do with the drop in sales. They responded that it was immaterial but would reverse all of the changes implemented by the new system and the methodology given another review to be on the safe side. Ouch. Not a stellar debut for their new data mining methodology. What went wrong? Was there a way to avoid this outcome?

This was a case of not getting the right people together to develop the solution. The company's data were complex, massive, and from various sources both internal and external. Sometimes, models alone cannot provide the answer. Some important aspects of the items were not directly part of the data. Examples include; upcoming discontinuations, where items are on their life cycle curve, what substitutes are available and stocked, and how items fit into the good-better-best strategy. People in the company had the knowledge but it just was not in their enterprise resource planning (ERP) and point of sales (POS) systems. If they had pooled all of this together – data plus subject matter experts (SMEs) - the scenario would have played out much differently.



This paper will provide some practical suggestions for leveraging subject matter experts. You may be the one with the mathematical and programming skills to do the data mining. However, the input of the SMEs will greatly complement your work. I hope that you can avoid your own disastrous analyst conference call. The presentation accompanying this paper will include additional examples and interactive portions not reflected here.

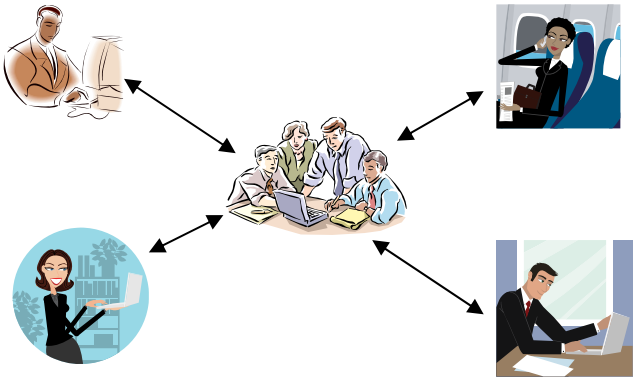
THE TEAM

One of my favorite questions to ask students was, "Is X a valid average for this variable?" Here, X represents some number I would write on the white board. Say I wrote 35. Some would answer, "Yes," right away. Some would ask what it represented. I would add that the name of the variable is, "age." More yeses followed. Then I would add that the data were for senior citizens receiving Medicare. Is this still a good average? The yeses turned to nos. The purpose of this class exercise was to get students in the habit of asking questions. Just getting an answer is not good enough. Additional information puts the number into context.

In the scenario above, a person somewhere at the fictional company would have known that the values in the data were ages of senior citizens receiving Medicare. If the modeler is unfamiliar with the details of the project, or worse yet just given data and told to model some target outcome, it is easy to see how errors occur. The person performing the analysis needs to seek out as much information about the data as possible. If people involved in the project had discussed the data in a meeting, you would walk out knowing what to expect as valid distributions and descriptive statistics for the variables in your data.

Suggestions for who to include in this meeting and who should be part of your data mining team will be discussed in more detail. The type and structure of the team will then follow. First off, we will review the role of the team. You want to know what the team is and is not going to contribute the project right off of the bat. Team members should know coming into the project what they will provide to the group. This foundation will help with the other aspects of your team.

One rule of thumb is to structure the team based on the type of project you are conducting. For projects that can end up on an analyst conference call, you would be wise to consider a larger and more formal team. Smaller and less formal groups may better serve exploratory projects or those that may have only a minor impact on the overall organization.



ROLE

As the analyst, you will be doing the data mining. It is not wise to make the actual modeling a group effort. The team should serve an advisory role and help with the understanding of the data. You bring the skills to build and select the models to the team. As you are forming the team, you want to share with them what you expect them to bring to the team and what you will be doing. Let them know before the meeting what you need and they can come better prepared to help.

At times people are reluctant to get involved with these types of projects since they feel they do not have the mathematical knowledge to do the work themselves. It is important to set the tone early on as you start forming the team. The goal should be to explain what the project will help accomplish and that you need their input to make it successful. This is not the time to impress people with your technical vocabulary and modeling prowess. Providing information on what you are modeling and why will help them better help you.

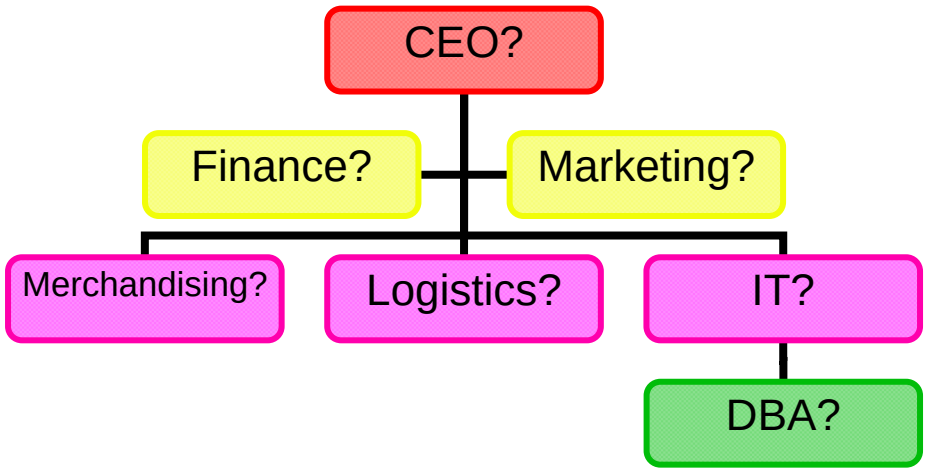
They do not need all of the details. For example, if you let them know you are working on a project to determine the factors that increase student retention using admissions data, financial aid data, and student record data they can come to the meeting prepared to discuss the subject and the data sources they are familiar with. Team members will be in a position to offer additional information as opposed to just answering questions. People may also bring information from current research on the topic that can guide you to additional data sources.

Formal vs. informal? I have worked on teams chartered by the President and on informal groups that I worked with on an ad hoc basis. Both can work equally well. Again, the real determinants are complexity and impact. The key is to get the entire group together in one place. There have been many times where I have learned something quite relevant based on the dialogue between other team members. Their perspective-generated questions were ones that I never would have thought to ask. In these instances, having everyone together in a formal meeting was advantageous. Scheduling these meetings could be the biggest challenge but you may not need many of them to accomplish your goals.

After the first meeting there are times when going to someone's desk when I need an answer is more appropriate. Sometimes you just need a short answer to keep you moving. Getting 10 people together in this case will only slow you down. Other times it may be helpful to get part of the team together. If you are only looking at a specific bit of pricing information, it may be sufficient to get the SME and someone from IT together.

MAKEUP

Where should you get your team? That answer depends on where you are getting the data.



Having the right information about the data is critical. Considering the volumes of data we deal with, it is unlikely that one person will have all of that information. We need help understanding what the data are and what is expected. Many times the team will include individuals from many parts of the organization. In the analyst call example, maybe category/brand managers would have all of the information. This, however, may not be a small number of people to work with. There may be many brand managers in the company.

Maybe you would need to include people from pricing to understand the strategy of each category and what type of dependencies exist (e.g., always stock a better if there is an item at the entry level price point for category Y). People from information technology (IT) who understand the structure of the tables may be of value. Possibly someone from the merchandising group who has a broader understanding of the interrelatedness of products would be useful: you may want to stock an entire "job" or "ensemble" as opposed to only some of the pieces.

A general rule of thumb is to include someone from IT. They have intimate knowledge of the data and its structure. As will be discussed in the data section, their knowledge base can assist you in many areas from access to content. I have attended a few presentations over the years where baking cookies for IT is encouraged. I am not one to promote bribery, but developing a true working relationship is important. Letting them know your level of database and data knowledge is also useful – this is not just a one-way street. It helps when they are comfortable that you know what you are talking about or asking for as opposed to just repeating jargon.

Size the team based on the number of subject areas and potential impact of the action resulting from implementing the results. Make sure there is at least one person who can answer questions for each area you are gathering data. If your project could have millions of dollars worth of impact to the company, you are better off erring on the larger side. Missing a single piece of information in something this size could have wide-ranging impacts. More focused projects might work well with a smaller team.

DATA: OVERVIEW AND ISSUES

Anyone who has programmed for any length of time has encountered dirty data and cryptic variable names. Even a variable with a name as simple as "cost" can be misleading. One time a new analyst ran a report to highlight items that sold below cost. The vice presidents were furious and ran to each team to see why they would do such a thing. After the panic settled and we had some time to review the report, it became clear that the analyst just used the cost variable compared to the retail variable. What they forgot to do was adjust the cost by the case quantity. Therefore, misrepresentation on the report happened for anything purchased in a pack greater than one. Again, a lack of understanding the data let them down.

Knowing the variable's name, type, and contents is not enough. You need to know what it represents. Age is one thing. The age of senior citizens on Medicare is something different. The larger context can make a great deal of difference. Subject matter experts have this knowledge since it is something they work with every day. Many of you may be familiar with the SAT college entrance exam. This is a data point I use frequently in my job. Much of my knowledge comes from when I took the exam but much has changed since then. The format of the test changed and in April 1995, the score scales were recentered. Our datasets have both the old and recentered values in one variable. What are the implications of using pre-1995 scores in an analysis with the new scores?

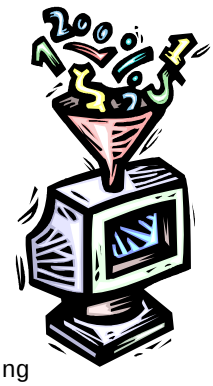
GATHERING

Sometimes just getting to the data is a challenge. It may be on multiple servers, in several formats, and contain hundreds of variables. Getting help to wade through each database may be a wise decision. You may need one SME for each database plus assistance for your IT group to help you in getting access to the data.

Data mining typically requires having all of the data in one place so you will need to pull the data from all of the sources and find a common repository. Then you have to combine it. Using student data as an example, you may have admission information with test scores and demographics, financial aid data, program of study information, and course level data. Hopefully you have a single variable, possibly student identification number (SID), to combine the sources. Even this simple example gets complicated. For many years, colleges used Social Security Number (SSN) as a SID. However, foreign students do not necessarily have an SSN at the time of application so temporary numbers are assigned. Once the student receives an actual SSN, the updating of the table occurs or maybe records will just start to have the new number so the database has multiple SIDs for one person. How can you identify and combine these records? Maybe a SME is the answer.

EVALUATING

After compiling the data, it is time to give it a quick look. Since data mining involves such large quantities of data, you cannot reasonably look at everything in depth. You need at least a cursory understanding of what you have and what it appears to represent. Maybe it is a review of the metadata or some simple frequencies. This will prepare you for asking the right questions.



If a value is missing from a single variable, data mining tools will typically exclude the entire record. Are null values representative of something or are they truly missing? In some cases, it may be desirable to assign a false value to keep a true record. One example of where this can occur is in survey data. Certain questions may only be answered when the preceding response is affirmative (e.g., if you exercise, how many days per week do you exercise). The response is not missing but rather not applicable. If the survey software codes these answers as missing, you need to know this and have a plan to address it. One particular project I worked on was evaluating data using a 1 to 5 Likert scale. The database had been redesigned and it seemed odd to me that everyone was doing better on their evaluations now (1 was "exceeds" on this scale). I looked at the data and saw many zeroes in the response fields. The old database contained nulls and the zeroes were now figuring in to the averages. This prompted a visit to the database administrators office and he informed me that he wanted to help out and fill in the missing data with zeroes in the view that I had access to. After thanking him for the thoughtfulness, I asked that they remain missing.

Other times, combining values is the answer. If a student does not have an SAT score, were they admitted based on an ACT score? These tests are on such different scales that putting the scores together into a single variable is a bad idea. How should a single, new score be created? Can an ACT score be converted to an SAT equivalent or vice versa? If yes, how and does my institution have an existing process to do this? You can see where this is going - right to the SME's desk.

It is also important to understand any outliers in the data. Are they true data points that add value to the model or are they mistakes? SMEs can help here are well. They can let you know if a data source was hand keyed (read this as prone to error) or generated electronically. You may be more willing to leave in low retail prices if they were pulled directly form the POS system (are workers giving excessive discounts?) than if they were hand keyed from store receipts by summer interns. In this case knowing the source makes the difference in your decision-making.

UNDERSTANDING

Knowing the relationship between variables is the next step. Is one just a multiple of another and therefore unnecessary? Have the contents of the variable changed over time as in the SAT example? This scenario is one that I encountered recently. At one time, the total score came from combining two sub-scores. These scores were part of the admissions decision process. For some reason, one of the sub-scores was discontinued a few years ago. The total score was no longer a total. This was important information to know. My models contained either the two sub-scores or the total to see if there was a difference in explanatory power. I learned of the change in usage during a meeting to discuss preliminary results with the SMEs.

Another aspect of understanding is the determination of nominal vs. ordinal. Does magnitude really matter? It may be intuitive for things like age, SAT score, gross profit, and retail sales. I have seen instances where race is stored as a numeric variable as opposed to a character variable. A race value of 5 is not 5 times greater than a value of 1. You may be able to spot this easily for a variable named simply race, but not necessarily for a variable named DW_ihs_student_rcode. Luckily, SAS makes it is easy to handle the issue either way. You just need to know how to handle it.

PRESENTING RESULTS

Prior to finishing your project and presenting the results to your superiors, you should present it to your team. It would also be a good idea to present interim and final results to your team. You do not need to make this a full-blown presentation where you stand in the front of the room. Again, you can choose how to best meet your needs and leverage the knowledge of your team. Including them shows the accomplishment from their input. Meetings for the final presentations tend to exclude database administrators, other analysts, and data stewards. This may be their only chance to see the outcome.

This extra step will serve two purposes other than showing the team that the effort produced results. First, it will give you a chance to test out your presentation. You can see how well it flows and determine whether people understand it. I cannot tell you how many times I have seen executive's eyes roll into the back of their heads when a presentation goes off on a technical tangent. You may rightly think the response curve is very interesting and telling, but more often than not, they just want to know how it affects the business and whether or not there is an associated cost. Your team can tell you how if flowed and how well the understood it. You can also pick up any non-verbal cues and tweak your presentation accordingly.

The second purpose for this dry run is to catch any errors in your understanding of the data. It is better for a team member to notice this than an executive. One small mistake or misstatement could derail your entire presentation. This was a hard lesson learned early on and a painful one to watch afterward. These pre-presentations have helped me better explain the results and to make sure I am using language that is consistent with other parts of the organization. Maybe the term is persistence. Admissions may define it one way and the Registrar another way. If both departments attend your presentation, they could walk away with very different interpretations of your presentation. With all of the terminology and jargon in use, it is easy to see how this can happen. Working with a team can identify these situations and allow you to decide ahead of time which terms need definition when you present results.

It is one thing to use complex technical language. It is another thing altogether when you are talking about one thing and everyone else thinks you are talking about something else. As previously noted, a pre-presentation meeting is how I learned about the change in a particular admissions data point. It was much better having that pointed out in the small group as opposed to when I was presenting to the Vice Provost.

CONCLUSION

It is no longer the case where we do not have enough data to do a data mining project. We can be easily overwhelmed with the sheer volume of data that is available. Knowledge about data tends to be dispersed throughout an organization. Getting information from the people who know the data and its structure, whether formally or informally, will greatly enhance your probability for success. This enhances the overall project and helps with buy-in to the results of the model. You will not be seen as an analyst working alone in a black box with no understanding of the business. The more you seek out the business knowledge contained in the data and through the SMEs, the more value you provide to the organization. Moreover, no one wants to go back and re-do an entire project after his or her CEO was questioned about it on a conference call.

CONTACT INFORMATION

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

Stephanie R. Thompson
Rochester Institute of Technology
Nine Lomb Memorial Drive
Rochester, NY 14623-5603
Work Phone: (585) 475-7237
Fax: (585) 475-7950
Email: Stephanie.Thompson@rit.edu
Web: http://finweb.rit.edu/irps/