**EMBA 8150 Data Analytics**

**The Initial Client Meeting**

Dr. Satish Nargundkar

The entire data mining / modeling project hinges on getting a clear understanding of the task that lies ahead. Asking the right questions at the initial meeting with the client (internal or external) can go a long way towards ensuring success of the project.

**Example of a project in the Financial Services Industry**

Assume that you are the lead analyst in charge of delivering predictive models to the client, a bank or finance company. The first meeting typically will involve your team and people that represent the client, typically a Marketing Vice President, a Risk Manager, and possibly an IS person.

The key topics that must be addressed during this meeting are as follows:

1. Define Project goals
2. Review Creditor’s Org. Structure, Policies
   * + Automatic Declines
     + Exclusions (employees, fraud, deceased, etc.)
3. Definition of the Dependent Variable, Outcome Period, Sample Time Frame.
4. Data Availability
5. Implementation Issues
6. Project Planning/Scheduling

**Project Goals**

While you will have a general idea as to what is needed, this is the time to clarify what exactly the customer wants. Typical goals for a bank or other financial institution include increasing response rate to mailings, increasing the approval rate (getting the right people to respond), reducing delinquency rate among customers of credit cards or other financial products, improving pricing strategies, decreasing customer churn.

Note that many of these goals could apply to any institution, not merely financial services. Utilities, for instance, could be dealing with all of the same issues. As an exercise, think about a completely different industry, say healthcare. What possible data mining goals might a large hospital have?

**Client’s Organizational Structure and Policies**

Organizations are always undergoing changes large and small, and it is useful to know what changes occurred in the client’s firm as it pertains to your project. Considering that you will be analyzing large amounts of their data, the project depends on the integrity of those data. Mergers with other companies create a host of problems for the client (and consequently for you) with the data. Typically, the company that merged with your client’s company will have had at least a few different variables in their database. The merged database may thus have a larger set of variables than the client had before the merger. However, some of them will only be populated before the merger period, others only after, and some will have data throughout.

A second and more insidious issue that muddies the data is that there are different definitions of the same variable. The two merged entities may have the same variable name with different meanings, or the same values within a variable with different meanings! In the financial services, it is possible that one company used credit reports from Equifax to populate their customer files, while another used Experian. While customer credit information should in essence be the same, there are many subtle differences in how different agencies report this information, and in what information is requested by the financial institutions. In the case of a merger, these disparate data now must be reconciled. As an example, take a simple variable like “Number of Trades” for a person (a trade is a financial contract, like a Visa card or an Auto Loan. If you have exactly one of each and nothing else, then you have 2 trades in your credit report.) Does this include those trades where you have a joint account with a spouse? What if one database included such accounts in the total count, while another did not?

Can you think of examples of potential data merger problems in other industries?

Apart from data integrity, it is important to know what is relevant to your project and what is not. Assume you are to build a model to help your client bank predict the risk of delinquency of a customer. The bank wishes to use this model to decide on the credit terms it will offer to its future customers, as well as to reject any applicants deemed too risky. What if the bank had a policy that its own employees would automatically be offered credit at a different rate than other applicants, and that they were guaranteed to get some credit? Clearly, the employees of that bank are not the potential targets for this model, and data regarding those employees should not be part of your model. There may also be policies that the Bank is considering implementing in the future. Suppose they have customers who got loans in the past despite a previous bankruptcy. Assume the bank now decides that in the future it will not grant credit to someone with a bankruptcy on their record. Since this is a policy decision, people with bankruptcies do not need to be in your model. In other words, they will automatically be rejected anyhow – your model’s job is to predict the riskiness of those who do not have a bankruptcy.

Consider admission to a University. What sort of model is used to make admission decisions? Are there any policies that might on occasion supersede the model?

**Defining key parameters – Dependent Variable, Outcome Period, Sample Time Frame**

One of the difficulties many beginners tend to have is with clearly visualizing what the data ought to look like - specifically, identifying the dependent and independent variables, the level of aggregation of the data needed, and the time frame from which the sample is analyzed. The client is looking to you to tell them what data you need. You must understand these ideas clearly if you are to ask the client for appropriate data.

**Dependent/Independent Variables**: If the goal is to predict the risk of delinquency in a potential customer, what data can be used to build the model? Obviously, one has to go back to data regarding existing customers, to study the differences in characteristics of those that were delinquent from those that were not. So a simple dependent variable might simply be the status of an existing customer – delinquent or not delinquent. If a two-valued (or more, but still categorical) dependent is used to classify existing customers, then a key question must be answered. How delinquent is delinquent? A typical client bank may have a response like “any customer that is 90 days or more past due is sufficiently delinquent to be put in the “bad” category. All else could be considered “good” or “not delinquent”, or one can have intermediate categories for those that are 60 days past due or 30 days past due, as opposed to those who are current in payments. So a dataset might look as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Y (Status) | X1 | X2 | X3 | X4 | X5 | … | … | … |
| 1 |  |  |  |  |  |  |  |  |
| 0 |  |  |  |  |  |  |  |  |
| 0 |  |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |  |
| 1 |  |  |  |  |  |  |  |  |
| … |  |  |  |  |  |  |  |  |

Where 1 = Good (Current in payments) Customer, and 0 = Bad (delinquent – 90+ days) Customer.

What of the independent variables? What sort of characteristics would you study to build a model to discriminate between the delinquent/non-delinquent customers? If the model is to be used for future applicants for credit, then the variables in the model should be those that you would have access to at the time of application. This is typically the person’s credit history as of the time of application, as well as any information in the application form itself. As discussed before, credit data includes information on Trades, Inquiries, and Public Records, while application data may have information on how long a person has been on the job, how long at current residence, income, and the like.

**Outcome Period:** Time is a vital element in understanding what data you need. Going back to the dependent variable, it is easy to identify the “Bad” customers from past data, since they have been sufficiently delinquent at some point. How about the good ones though? Just because a customer has not yet been delinquent enough is not evidence that he/she will not be delinquent in the future. What if a person became a customer only 4 months ago, and has not become delinquent yet. Is that sufficient to mark him as a “Good” customer? Probably not – one has to observe someone long enough to judge. Of course one cannot wait a lifetime before judging a customer to be “Good”. There is usually a sufficient amount of time of “good behavior” that would lead one to categorize the customer as “Good”. This can vary by client, and is something you must ask about. It is in this case the Outcome Period. So an Outcome Period is the duration for which a customer must be observed to determine his/her status. A possible Outcome Period for a credit card holder may be 12 months. This means that no customer who has been with the bank for less than 12 months can be judged as a “Good” customer. They could possibly be judged as “Bad”.

There are two implications of the outcome period – the first is that your sample should only contain customers that have been with the bank at least 12 months (or whatever the Outcome Period is decided on.) The second implication is that your model will predict as far into the future as the outcome period. Once you develop the model, for instance, you will be able for a future applicant to assess the likelihood of delinquency in the next 12 months. This is the other approach to thinking about the Outcome Period. How far into the future do you wish to predict? If you wish to predict the likelihood of food poisoning within 24 hours of eating some type of mushrooms, then your Outcome Period is 24 hours. Assuming you are pulling data at this moment from a system regarding this, you would then avoid having in your sample anyone that ate those mushrooms less than 24 hours ago, since not enough time has passed for the outcome to be known as of now.

**Sample Time Frame**: Understanding the outcome period is thus crucial to determining the period from which one would pull the data. The Sample Time Frame is the duration during which, if a person became a customer of the bank, would be eligible to be in your sample. Assuming a 12 month Outcome Period and a start date of 1/1/2012 for the project, you cannot have anyone in the sample that became a customer over the past 12 months, since you would not have a value for the dependent variable for that person (except for those that already went “Bad” – you can pick those if you want in the sample). So assuming we pick people that became customers before 1/1/2011, how far back do we go? Do we include in our sample people that became customers, say, 30 years ago? That depends on what you consider to be relevant data. You want a sample that is not so old as to be irrelevant to current customer behaviors, while at the same time have enough information to account for seasonal variations. So a sample time frame may be something like 2 or 3 years in the financial services industry.

One can visualize the timeline as follows:

12-Month

Sample Time Frame Outcome Period

1/1/09 1/1/10 1/1/11 Today (say 1/1/12)

The sample will therefore consist of those people that became customers of the bank between 1/1/09 and 1/1/11. Remember that the only restriction regarding time is that they became customers during this period. The value of the dependent variable would be determined after observing the customer for the next 12 months. Thus, for a customer that joined on 12/15/10 (within sample time frame), payment data must be observed until 12/15/11 to determine whether he/she is a “Good” customer. In other words, data on a certain variable (or variables) during the outcome period are still necessary to obtain.

As for the independent variables (Xs), remember that even if a customer joined on 12/15/10, their credit history information may include their behavior for all their adult life. For instance, a variable in their history might be “Number of Times Ever 30 days past due”. This variable may count the instances in this person’s history for the past 25 years. In other words, the 2 year Sample Time Frame does not restrict us from looking at any amount of history for those customers. It simply refers to the **time frame during which the person was accepted as a customer** at this particular bank for this particular product.

If you are clear about the above concepts, you can discuss it with the client ask for the data that you want without any confusion. Note that you can ask for the behavior data for each person in the sample and create the dependent variable yourself (recommended) or ask the client to define one for you when they send the data. This would require them to do some programming and for you to provide very specific instructions.

**Data Issues: Availability Medium and Format**

Two issues that must be clarified regarding data at the beginning of a project are the availability of data and the way it is to be delivered from the client to you.

We can create a wonderful sample design with a certain sample time frame and a list of variables to analyze, only to discover later that data during that time period for some or all of those variables is simply not available for various reasons. It may be archived in a way that is too difficult or expensive or time consuming to get to, and the client does not want to do so. Knowing how the client archives their data, the time period for which they keep data readily accessible are all important questions to ask.

An even more trivial sounding but important question to ask, assuming that data are available, is the format in which they are available. You should specify to the client how you want the data, including formats and media for delivery. An analytics company that worked with PC based software received data from a client on large computer tapes used on mainframes, and simply did not have the hardware to read them! Yet another client provided data on media readable on PCs, but looked like gibberish, because they were still in the EBCDIC format used on mainframes, rather than the ASCII format used on PCs. In both cases, conversions of data to the right medium or format cost several days of delay in the project.

**Model Implementation**

Once the data are received, the analysis completed and a model built, how is the model to be implemented in the client’s organization? We may be building a risk model for a bank, which may be used in conjunction with a profit model and a response model by the company. If so, it may help us to know that. Are there elements of the other models that might interfere with this one? Would it be better to rebuild all three models together to best predict what the client wants?

Sometimes variables that are available in the historical sample may be ones that the client no longer stores or uses. If so, should you use them to build the model? What if your model finds those variables significant? Going forward, given that the client does not have them, your model cannot be implemented as specified.

There are sometimes other system constraints. An analytics company built a model that scored customers based on their risk, with scores ranging from 0 to 1400. The higher the score, the lower the risk of the customer, according to this model. As it turned out, on the client’s mainframe computer where the model was implemented, only 3 digits were assigned to hold the score variable, since the previous model they used only went from 0 to 999. The result was that when the computer computed a person’s score, if it was 1000 or above, it simply truncated it, and reported the first 3 digits! So a person with a score of 1265 (a potentially very low risk customer) showed up as having scored only 126, resulting in the company rejecting many potentially good customers until the problem was discovered.

**Project Scheduling**

A typical source of trouble in a project is conflicting requests from different sources within the client’s organization. As discussed before, it is useful to understand the client’s organizational structure, specifically for clarifying who will be authorized to make any changes to the contract with you, and who you will be reporting progress to.

The key elements of project management are making sure that cost, quality, responsibilities, and time elements are clearly agreed upon between the client and the analytics team. A table like the one below can be constructed at the meeting to ensure that the time and responsibility elements are addressed:

The times in the table are to be measured starting today (xx/xx/xxxx).

|  |  |  |  |
| --- | --- | --- | --- |
| **Activity** | **Client Responsible** | **Analyst Responsible** | **Duration for completion** |
| Sample Design - data requirements |  |  | 2 days |
| Getting data from archives |  |  | 7 days |
| Preliminary Report – Clarification of Data Issues |  |  | 14 days |
| Final Report with Monitoring Report Templates |  |  | 7 days |
| Implementation, Monitoring |  |  |  |
| Monitoring Reports |  |  | 3 months |

Using durations for activities rather than actual dates ensures that each period begins when the previous one has ended. So if the client delays the data and takes 3 weeks instead of 7 days, you cannot be held responsible for the original deadlines and still have 14 days for the preliminary report after you receive the data.

It is important to clarify also if the preliminary and final reports are to be made online or face to face, and if the latter, where the meeting is to take place – the client and analytics team could be in different cities, even different countries. The number of people attending also impacts the cost, and some agreement about the cost boundaries is useful to avoid future conflict.

**Summary:** The initial client meeting is a critical time for the analyst to ensure the success of the project going forward, by asking the right questions. The six key topics discussed above are essential elements of this meeting.