Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:05
Lecture:17

ANALYTICS PROCESS | Business Intelligence & Analytics

Hello and welcome back to this session on Analytics Process, as a part of business intelligence and analytics course. So we have laid a foundation by now for analytics as a practice in an organization. So alongside understanding how to do analytics, we are also trying to understand how analytics can be integrated with an organization. So essentially when you grow as engineers to managers or analyst to managers and you know, decision makers in an organization for analytics practice, you also need to think about the organization and how it can take decisions on analytics itself.

You know analytics as a separate sub organization that needs to be built and its culture needs to be built and resources needs to be developed etc. So by looking at the architecture and infrastructure from a technology standpoint we understood what are the requirements for starting analytics in an organization. And alongside we also understood or we try to sort of develop understanding about data and its culture in an organization, a data based culture in the sense how you take care of your data- how you segregate, how do you filter, how do you sort of create relevant data sets for future analytic purpose and store them separately and then what kind of interfaces you develop to generate insights from them.

So we are very much thinking from the organizational perspective at the same time developing understanding about what is data and data for analytics. So it is on data that we spend a lot of time now and along with data in fact we also learned how to do descriptive analytics. How to actually work with data that is stored and bring or pull out relevant data based on what decision makers ask and present that in the form of a list, how to run queries and how to run multi dimensional queries and what goes into it and so on and so forth. So that is a foundational aspect and I say it is foundational because without data and data as a foundation and data as a culture, it is very difficult to build analytics practice in an organization.

So now we get into that aspect of what is analytics and what are the steps involved in analytics. So we are now looking at analytics and data mining as an advanced form of analytics, from descriptive we are moving to explanatory and predictive analytics. That is, so it becomes more advanced. So I can also put a prefix, advanced analytics. Advanced analytics process and here we say that there is a process we underlie the term process.

Why we should worry about process in organizations? Again I am saying this is tied to organization because organizations go for process certifications. You must have heard about the ISO certification and similar kind of certifications that are available in the industry. You must have heard about capability maturity models. So it is again process certification and organization and its processes are sort of taken care or standardized and get certified for the standards they follow.

Why should someone be concerned about a process and why is process important ? Generally in a production or organization setting and also in analytics we are bringing the same concept or same principle to analytics as a practice. Process is important because process quality determines product quality. For example, if you take a bottle of water, bottled drinking water. One thing that we want to see is whether it is a genuine product or not and how do we know. Of course, now a days we look at the brand name. The brand itself builds certain confidence in us and then we also look at, is it certified by certain standard organizations?

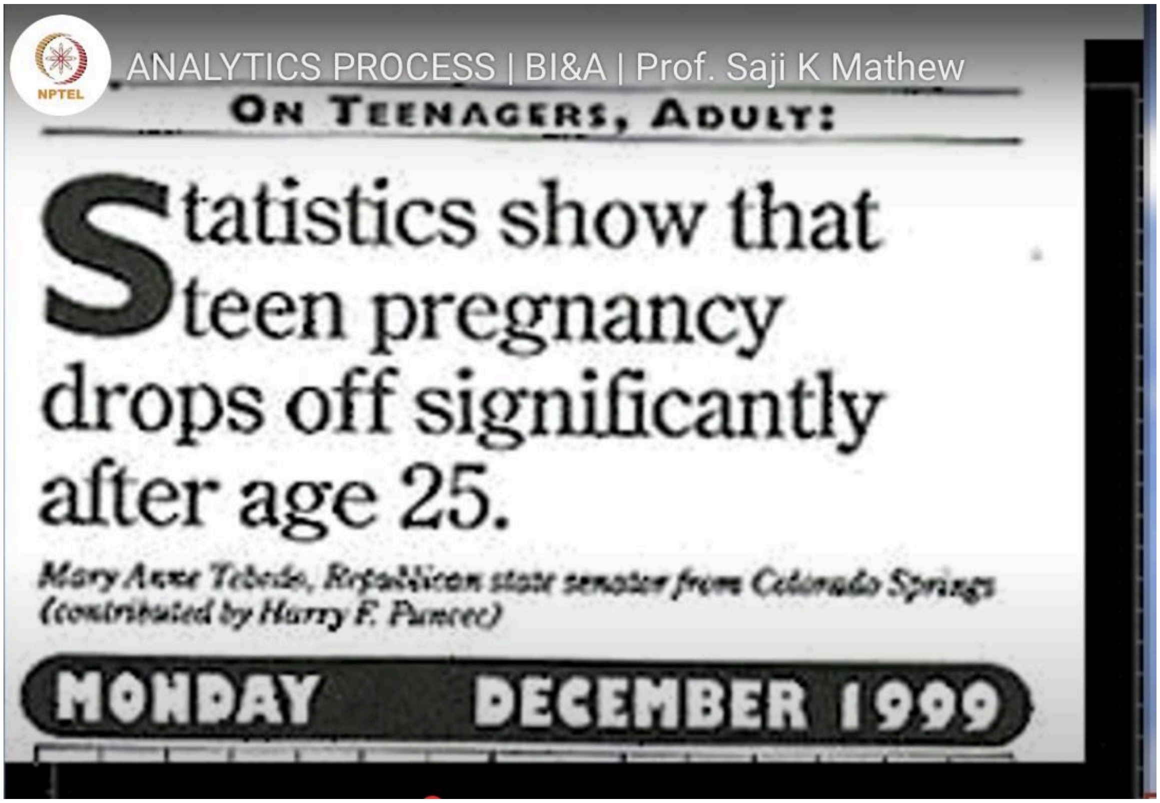
So we also can see some companies advertise about their ISO certifications. What do they basically mean? They essentially mean that in the production of this product, we have followed a standard process. We have followed certain standard process. Standard means tested, reliable, credible processes that have been followed in the making of a product, gives us the confidence that the final product is credible.

So when we translate this idea into analytics we are essentially meaning that if you are doing analytics ensure that you do it in the proper way. You follow a process for that. For example, if somebody gives you a problem and a data set. How do you go forward? For example, suppose there is no clear definition of problem. Somebody in willy-nilly articulated some problem and gave a data set.

If you do not follow a process and if you are early in your career, chances are that you look at the data and then you try apply all the algorithms or all techniques you are good at. And then create some outputs and report the outputs that you obtain by analyzing the data. I am not against it. It is good. You are learning a lot by analyzing data.

You learn something. That is good for you. That is good for you. But the purpose of analytics is to support decision making by decision makers of organizations. Business organizations, could be business organization or any organization but they have priorities. They have their priorities, their objectives.

Analytics should help meet an organization reach its objectives. Not reach your personal objectives. So therefore it is very important that instead of applying what we know into the data, we rather follow a process in analytics. And that is the reason why we are talking about analytics process.



ANALYTICS PROCESS | BI&A | Prof. Saji K Mathew

ON TEENAGERS, ADULT:

Statistics show that teen pregnancy drops off significantly after age 25.

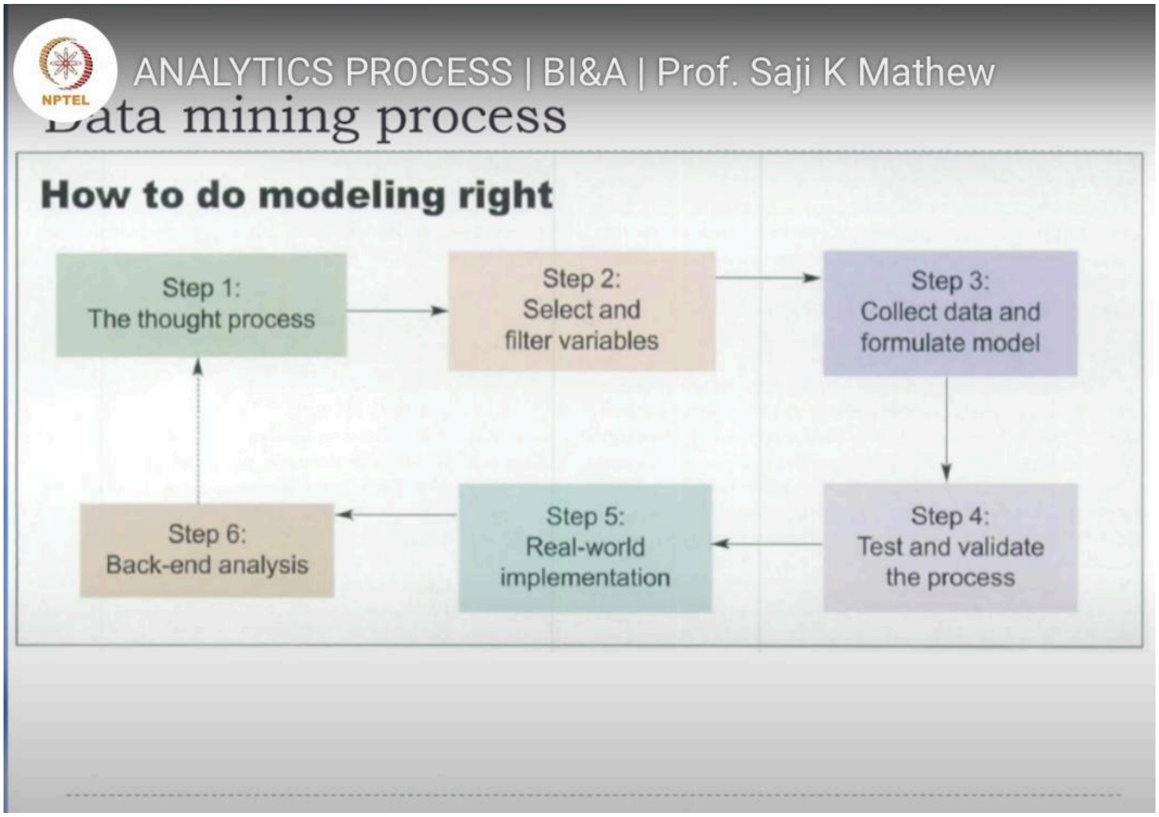
Mary Anne Tebedo, Republican state senator from Colorado Springs (contributed by Harry F. Puncec)

MONDAY DECEMBER 1999

Let us move on. So as I said the presence of an analyst along with his mind, along with the sound mind is very important when you work with analytics problems and data. Otherwise you may reach very funny and laughable conclusions like this. If you use a software or a tool for data analysis, it will, so long as it has received the right data, it will throw out some results but it does not mean that it is a useful result. So let me stress on mindfulness.

Mindfulness is a property or it is a characteristic of a individual or a person. So what is a characteristic that an analyst should have before getting into analytics? One should apply one's mind. One's mind should be active to ask a lot of questions, a lot of relevant questions before one takes decisions about anything. We will soon see that analytics would follow a process meaning steps, criteria etc. That is what a process consist of. So therefore in each step one needs to stop and think. And if you do not do that, again you can see certain funny inferences that is drawn from data.

So mindfulness, being present, being actively present and being inquisitive, being doubtful often. Ask questions, doubt data, doubt credibility of data before you start using the data. So there are lot of properties that as an analyst you need to have before you get into analytics. So let me present to you since I promised to talk about process. I have used data mining and analytics synonymously here, although they may not be the same but the purpose of both is to use advanced techniques to provide insights for decision making.



ANALYTICS PROCESS | BI&A | Prof. Saji K Mathew

Data mining process

How to do modeling right

Step 1: The thought process

Step 2: Select and filter variables

Step 3: Collect data and formulate model

Step 4: Test and validate the process

Step 5: Real-world implementation

Step 6: Back-end analysis

So from that point of view what are the steps involved in analytics? So I have taken this lesson from Kumar and others work. He is an alumnus of IIT Madras and there is an interesting paper that is authored by Kumar and others and they have, based on their consulting experience reflected on the work they did and provided certain approach to work on analytics. And the result is a six step process and the steps are number one, thought process, number two, select and filter variables, number three, collect data and formulate model. See that this is sequential, one after the other. And fourth step, test and validate the process. I would also cross, I will also use the term model. You built a model there, you need to test and validate the model. And fifth step, real world implementation. So so long as, up till here it is laboratory and here you are going again to the field or to the real shop floor or organizational setting where you have to apply your model, real world implementation. And then when you implement a model, you will find that you, the results from the application of the model may be very satisfactory, satisfactory, less than satisfactory or it, you know it even worsened your results.

So therefore it is very important to look at, here you look at business performance. As a result has the business performance improved or deteriorated and if improved to what extent? You need to have some measures also for that. We will see all that as we go. How business performance improvement can be measured as a result of analytics.

And therefore most of the time, since nothing that we make will be perfect in the first iteration, first run, you may have to make changes. You may have to change your thinking, you may have to include or remove variables, you may have to use a different model or you know, you may have to collect more data. Lot of lessons will be learned when you implement a model. And so therefore, it goes through iteration. So it is not ending, it is a cyclic process, you can see that.

So that those are the steps involved in analytics. Now I do not have a slide to explain to you what is thought process. Before I go to the next slide we need to think about what is thought process. What is thought process? Thought process cannot be taugh, cannot be taught. Thought process need to be learned.

So it is something that you learn through doing, learn through thinking, learn through reflections. One example that I can highlight here is from the case that we discussed on bizocity scoring at AT&T long distance. We noticed that the organization or a particular department of the organization had a problem and that was a real problem that they were not able to target and build their customer base. So they have a problem. Now they were not able to solve that problem internally.

They did not have the analytics capability inside. So they employ an external organization. Of course, it is a part of the group but it is external to that particular organization. So the consultants come in.

Now that is a point to reflect. What did they do? What was their thought process? That is where, before they went on to develop some logit model, there was a thought process involved. There is a, in other words, they developed an approach or they developed a methodology. In research language, we use the term methodology. Before you do research, you should have a methodology to do research or in a similar terms before you move on to analytics, you should develop a proper approach to doing analytics. And that is the thought process involved.

How are we going to address this problem using data? How are we going to address this problem using data? So therefore we can also call this particular stage or phase as business problem to analytics problem translation. Remember in your course outline, one of the objectives is how do you translate business problems or problems which are general in nature to more specific analytics questions or analytics problem. Business problems can be described as we did in the case of bizocity score. We can also put business problems in the form of questions.

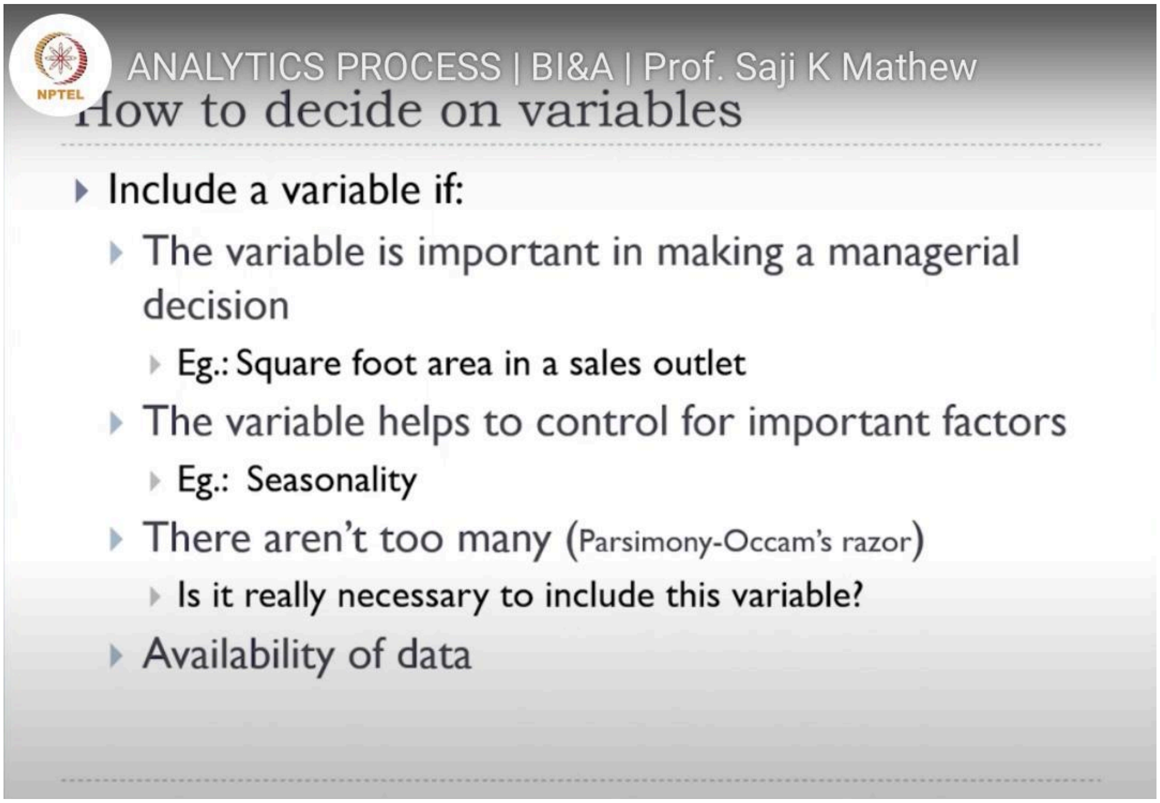
And those questions could be translated into analytics questions. So in hypothesis testing, in the world of hypothesis testing some of you must be aware, there are actually general hypothesis that you write which as decision makers you want to test. And those hypothesis are converted to statistical hypothesis, meaning hypothesis which can be tested through statistical techniques. That is when you have to be very crisp or specific in terms of defining your variables and defining the relationship between variables. So in a similar way, the thought process converts a business problem into analytics problem.

So the consultants in bizocity score thought through the problem, describe the problem and then they found there is a twofold issue. One is data source itself, other is how do you classify prospects into business and residence. And the current process was very unreliable. They developed a method. So that is where actually they started thinking.

They started moving from business to analytics and started saying, well if we build a model of this kind and if you score each prospect with number between 0 and 1, then the decision maker can know whom to target and whom not to target. So they are thinking from the decision maker'sperspective. They are also thinking from data perspective. They are also thinking from what are the relevant variables. So there is a continuum between the thought process and step 2 which is about selecting and filtering variables.

In the sense your choice of variables depends on what is the technique, what is the method you have in mind. Here in this case score, score each object, that can help the decision maker to take decision. So therefore you look for a scoring model like the logit model which will give a probability score to records or objects. So therefore that is the thought process they followed. Look at from decision makers chair, look at from data angle, look at from technique angle, what method can be used and so on.

So now by the end of thought process, you should be clear about what is the final output that you are going to give to the decision maker, what is the output that you are going to give to the decision maker. And then to create that output, what are the algorithms or techniques that are prevalent or that can be used. So you have to develop this understanding and only then you can move on to the select and filter variable stage.



ANALYTICS PROCESS | BI&A | Prof. Saji K Mathew

How to decide on variables

- Include a variable if:
  - The variable is important in making a managerial decision
    - Eg.: Square foot area in a sales outlet
  - The variable helps to control for important factors
    - Eg.: Seasonality
  - There aren't too many (Parsimony-Occam's razor)
    - Is it really necessary to include this variable?
  - Availability of data

So let us move on. So we come to the next topic which is about deciding on variables. This slide provides you four bullets or four points to keep in mind while selecting variables for a model. Ultimately in order to solve a problem using advanced analytics, you need to have some model, some method, some I would say analytical method be it an algorithm, be it a statistical technique you have to think of something. Only then you can actually decide on variables or that is important. There may be going forth and back when you do all this but it is important to have, well I am going to use a logit or a decision tree or some model which will actually give me a probability score at the output.

Then I need to have a set of variables to build that model, to develop that model and what are those variables. What are the criteria? Criteria one is, include a variable if the variable is important in making a managerial decision. So keep in mind in most of the techniques that we are going to learn, not all of the techniques but many of the techniques, there will be a y equals fx relationship where y is the outcome variable and x is the explanatory or the x drives y. The variance in x and the variance in y, they are moving together, they are correlated, there is association between y and x.

So x determines y, x is a determinant of y. So we try to find out the x variables that influence y. That is the way we model a lot of statistical and algorithmic models. There are others as well, we will come to that but let us take for example, your model what you have in mind is something like this. For example, a good example is a regression model or a simple linear regression which many of you must be aware of.

In which case there is a y and there is x. This is the data that you have in hand. This is the data that you gather or collect. So simply take an example of what determines the sales of a retail company. Sales is a function of square foot area, the marketing mix model.

So it depends on certain well-known variables. One of them is the square foot area, price, promotion and so on. So promotion and other variables. These are the x variables and this is the y variable. Now how do you know this? Suppose somebody gives you a problem, determine an appropriate location for situating a store. It is a store chain, the store already has 100 outlets already.

But you have to find out 101st store location. So the simple, when you try working on this problem, it will be a fairly complex problem. But broadly speaking, what is a suitable location for a store? One of them is, of course it should be a location which maximizes sales or you get a lot of revenues. There are other factors also like the environment and legitimacy and all that in organizational theory. But if you go by the pure profit motivation, you try to maximize sales. So therefore you look for those variables which are relevant to modeling sales as a function of the x variables.

So these variables become important. But where do you get this information from? How will you find out, which are the relevant variables? Relevant variables, again not all problems are the same, not all contexts are the same. Suppose you are given a problem which you know as an analyst you are not familiar with,what is your first step that you do? You can enumerate that. One, experts or subject matter experts, those who know about the problem, go talk to them. In large analytics organizations, they will have subject matter experts who will inform the analytics consultants or they work together, so that you do not make any mistakes.

This expert knowledge is one source and other is, a lot of knowledge is there in literature. We call it extant literature here. I will call it extant in the sense all kinds of literature including published books, published research papers, even white papers where, well you have information about this kind of problems which you are addressing. If you do not read for example research papers where problems have been well described, solution approaches have been well described, tested results have been provided, if you do not care about that, essentially you are not approaching your problem scientifically.

Not all problems need to be discovered from the scratch. Problems already exist, solution has already been discussed and documented in books and papers. Therefore, a good analyst who has a scientific approach should read literature, should have access to relevant research literature. So Google Scholar is something that anyone can access today, free. And when you access that, you find you get a lot of reading free. Of course if you are a research scholar then your job is to look for other sources of research papers.

But as an analyst working in industry at a minimum level, one should be able to go, describe your problem in the form of keywords to a Google Scholar and get some good papers and read them, to start with. I would suggest that the sequence be the opposite. Before you go to an expert, develop understanding about the problem. And only, otherwise the experts may not entertain you, you know nothing about it. So I cannot teach you from A B C D, learn A B C D and then come. So this is first step. I already called it first step.

The second step is the if the way include a variable, if it helps to control for important factors. So we call them control variables. Look at a model like this. Sales is equal to square foot promotion. Well, these variables are important relevant and they are also something which the management has control on or management can decide on. But the literature is suggesting us or the suggestion here is include those variables which also will influence the outcome, even if you are not able to change it. A good example is seasonality.

You may add seasonality as another variable. Seasonality is not something that you are interested to study or the management cannot change seasons. But this needs to be in the model, otherwise the model will not have enough explanatory power. The model will not explain or the model built, if it does not have all the relevant variables, will not give a good picture. And that aspect we will see when we work on one problem like adding one variable to a model, changes the strength of other variables also in the model or the coefficient of other models, other variables also in the model. Therefore it is important to include all relevant variables which has significant impact on the outcome. So that is another suggestion. So just because seasonality I cannot control seasonality why should I bother about it. That is not the point. If seasonality is an important determinant of an outcome, include that also so that your model overall has explanatory power. That is the second principle.

The third principle is, there are in too many. That is a very interesting principle and that we need to understand well and that is a principle we will be applying when we work on certain problems in near future. So I will explain it in a little more detail in a, soon.

Before that let us look at the fourth principle which is availability of data which is like a common sense. For example in the bizocity score, the best way the most intuitive way to understand if a customer is business customer or residence customer is to listen to the conversation. Instead of looking at whom the customer is calling, what time he or she is calling etc. If you have the voice, it is much more easy. You can actually do the use the audio data. But unfortunately that data is not available. You may try to include a variable but data is not available. There is no point. You have to look for alternate variables or proxies or variables which are related to that variable, for which data is not available.

But data is available for, to inform the outcome in different way like, as in the case of bizocity score, the who is calling whom and what time, sort of data which can inform about the nature of the customer. So you look for alternates, you look for proxies, you look for other indicators. If data is not available, don't include a variable in the model. But don't have too many variables is a principle in model building.

We technically, it is called parsimonious models. Variables which have just enough variables that is required, not more than that. So Occam's razor is a broad philosophy which is about to solve a problem using minimum approach, only have minimum things, don't have too much. And parsimony actually means the same thing. Don't have too many variables. Thank you.