Relevance of Analytics in the Context of Decision Making – Part 1

Prof. Ram Gopal

Profile:

My name is Ram Gopal, and I've been a professor for the past 25 years most of it in the U.S. even though I've had visiting stands both in India and in China and particular for the past one year I've been affiliated with IIM Udaipur as a visiting faculty, and in the past year I have had the great pleasure of interacting closely with Director Janath Shah and as well of worked on several projects. I'm currently working on the number of projects with some of the bright young faculty at IIM Udaipur. In terms of my research and my teaching interests, they are in the fields of information systems and in particular analytics. That's where I've done a lot of my work off late and also in operations.

In terms of my research projects over the past two, two-and-a-half decades worked on issues around how information technology and infusion of data digitization has impacted industries, variety of industries. I've worked on issues around technology infusion in the entertainment industry, in the healthcare sector. I'm currently working on FinTech, mostly how is information technology impacting financial activities and the work that I will present today is more about the use of analytics in the context of brick-and-mortar retail companies.

I have organized the webinar into two parts. The first part is what I call a primer on predictable and prescriptive analytics. The second part is on to my paper which is on predictive and prescriptive analytics.

About Analytics.

So, what is analytics about? Fundamentally analytics is about harnessing the power of data. So, analytics is all about data. The reason analytics is so important is because it's emerged in the last decade especially as perhaps the most important organizational asset. In the past, if you go back let's say two decades if you talk about what are the most important assets in a typical

organization it always used to be human resources and typically data as a resource would come in as second. But now with all the changes that you see in terms of machine learning AI, Big Data increasingly for most organizations data has emerged as the most important organizational asset and analytics is all about how do you manage the data and how do you harness the data and how do you create business value from the data. So, in a sense, analytics revolves around data.

Now, the study of analytics entails looking at the entire lifecycle of data; from birth to death so to speak. So, there are three broad topics that are of interest. One is data management which essentially is how does data get created, how is data born, what are your business processes, what are your business organizational activities that result in the creation of that data. Then how do you then store and manage it?

So, the first element aspect of analytics is data management. In fact the point I should make is that if you take a look at especially in the U.S. look at the demand for jobs in analytics lot of them really require data management skills. And so data management is, in fact, the critical aspect of analytics. So, if you have sound data management skills, if you have sound SQL skills then you are easily employable in the field of analytics. The first aspect of analytics is data management.

Now once you have data in, the next piece that comes into place is called predictive analytics. People often used this term 'data mining'. Machine learning is another term that's also used.

The idea of **predictive** analytics is to learn from it. So, predictive analytics is all about learning from data. As more and more of the activities we engage in are of economic or social kinds - what we do online - things that we do on social media, things that we do with our mobile devices; all of those activities and actions that are getting converted to digital data. With access to the data, there are lots and lots of exciting, interesting

important opportunities to learn. So that's why predictive analytics has become really important.

Then comes what is called **prescriptive** analytics. Having learned from the data, how do you make better decisions; how do you change certain processes in your organization to create value from analytics. So, there are three elements to analytics - the creation of data, learning from data and then making effective decisions that change your organization ability to reap the benefits of analytics. Predictive analytics is all about discovering knowledge, learning from data, and in terms of a more formal definition, essentially is, extraction of knowledge which is non-trivial, implicit, previously unknown and potentially useful patterns from data.

Predictive analytics broadly fall into two classes of models; patent discovery and predictive models. Now one way to understand the difference between these different types of predictive analytics models is, patent discovery is also called unsupervised learning whereas, predictive models are called supervised learning. The difference is that with unsupervised learning or pattern discovery, there is no specific objective that an organization has. You have access to large volumes of interesting data and what you would like to do is to sift through the data to see if there is anything of interest in there. With predictive models or what's called supervised learning, there's a clear business objective. You are looking to understand something and looking to predict something. So, with that goal in mind, we develop models to help you learn that with two different classes of algorithms.

Some of you may have heard of this very classic example of unsupervised learning. This is called the beer and diapers story and this is slightly dated now and the story goes that someone was looking through sales of different items in a convenience store. Convenient stores are places where people can just go buy something very quickly and doing unsupervised learning on the data what they found was that there's a high correlation between sales of beer and diapers. This is an odd combination that you would expect and that was

an interesting pattern that they discovered and then having discovered that they tried to understand why is it happening.

One plausible reason why that happens is, typically on week days/nights the father of a young baby has to run to a store quickly to buy diapers for the baby and while he's buying diapers he also picks up a six-pack of beer. So, that's an example of unsupervised learning where you learn something of value from the data.

Predictive modeling, on the other hand, has a very clear objective. There is a variable, there's an outcome that you're looking to predict and you would like to develop models to be able to predict the outcome basis the variable. Here is a very simple contrived example: think of a tax collection agency. In the U.S. it's called IRS Internal Revenue Service, and this collected tax collection agency might be interested in finding out who is who of the tax filers who are likely to cheat on the taxes so that they can be audited. What they would like to do is to learn about who is likely to cheat, and that's a question mark. With access to some data about individuals - whether they filed for a refund or not, their marital status and their taxable income - how can a predictive model predict if someone is likely to cheat on taxes.

Now if you think how you learn; given this setting what kind of logic would you use to predict who's likely to cheat on the taxes. And the answer is, by looking at the past data.

Now if you go back to the past, what you would have is information about all the variables, including the outcome. So, you need to look at your data from the past and see who in the past has cheated the taxes and who did not. You get the data from the past and you use to learn and apply it to the future. So that's the core aspect to predictive modeling. You need to have access to data from the past to engage in predictive analytics.

A little bit of terminology here. In this example I have three variables that I used to predict an outcome of interest. The outcome of interest is my output or my dependent variable or the target. That's an object of interest. Then my predictors are the variables that are used to predict are called my access, my predictors and my input variables. In this example my predictors are refund, marital status, taxable income, and my objective of prediction is to predict who's likely to cheat on the taxes. This is a very simple example just going to quickly run through several other; probably it's slightly more realistic examples and again many of these examples come from real settings and some come from, there is a site called Kegel [PH] which runs a lot of these predictive modeling competitions. So some other data comes from there as well. So here's one simple example where the object of interest is to predict delinquency. So, you have a number of individuals that you've known to and you would like to predict which of them is likely to go delinquent and not repair. So my Y here is who's going to go delinquent in two years. My predictors, I have a number of predictors here including the individual's age, the monthly income, debt ratio, number of dependents and a whole host of other variables. Here my objective is to predict who is going to go delinquent in two years based on a set of predictors. Here's an auto insurance prediction where what you would like to do is to predict which of your insureds is likely to get into an automobile accident then file for a claim. So, you would like to predict not only which of your customers is going to file a claim but also what the claim amount might be. In the predictors you have access to, here is information about the vehicle, about the individual including things like the gender, area they live in and the age.

This last example here this is an interesting one and this is based on a company called Carvana. Carvana is a company that sells used cars. Essentially what they do is they buy cars from several different auctions. They buy these cars, they work on the cars, they repair them, they improve them and then they resell. Typically Carvana earns net profits of about $1,000to $1,500 per car. So they go to these auctions and they bid on cars, buy them, work on them and then resell them. Now one

problem that Carvana had was that when they go to buy the cars from these different auctions, they found that about 10% of the cars they buy tend to be what they call kicks or bad buys. They are so bad that working on them and trying to resell them is simply not worth it because the car is too far gone or the repairs are too expensive that they cannot earn any profits. So, it happens about 10% of the time. Their objective here was before we actually buy these cars can we predict which of the cars that we are going to bid on is likely to be a kick or a bad buy. So their objective for prediction is this a bad buy or is this a kick what they call and the predictors they have are the whole host of predictors. Information about, variety of information about the car.

So these are a few examples of a typical predictive analytics project. Now from a business perspective before you start going down this path of predictive analytics there are a few questions that are really important to address. Let me just walk you through some of these questions. The very first question that comes up is why do you want to predict. Before you get into any analytics project, you have to have a clear understanding of what your business imperative is.

Now what goes with the business imperative is what is called intervention. Suppose you're able to predict whether a particular car you want to buy is going to be a kick or, which of your customers is likely to get into an accident… So what? You have to have some form of intervention. Before that negative or positive outcome happens you have to have a plan for intervention. In the case of Carvana, if you predict that this is a bad car then you don't bid on it. In the case of a customer likely to be getting into an accident, either you don't give them insurance or you change the terms of the insurance. You charge them higher premium, lower deductibles a whole host of different options. So you need to have a clear sense of what is the business objective that's driving their analytics. So you need to start with the why question.

The second thing is ‘when’. When do you want to predict? Obviously, you want to predict before the

outcome happens. So, there's no point in predicting if a customer is going to get into an accident after they get into the accident. It's too late! So, the timing of prediction has to be obviously before, before a claim is filed for example. But before can be at different points in time. In this example, you see the claim is filed. If you go back in time there's when the customer got insurance, you do the prediction, your choices are either to give them insurance or not or, change the terms of insurance. You can also predict after they become the customer, but then your intervention choices change to giving them more educational materials. You warn them about driving carefully. The question of 'when' is important because when you predict has two impacts; one is what you can do with it and second is what information you have. For example, if you try to predict before someone becomes a customer then the type of information you have about the customer is more limited but if you wait until they become a customer then you have more data, more data about what the driving habits were, whether they pay on time things like that. So, the different points in time become important, when becomes important.

In some organizations they have a series of models and each of these models is implemented at different points in time. So that's another thing to always keep in mind. We've talked about the why question, the when question.

The next important question is 'where does my access come from?'. These are my predictors. This is an important question to ask for several reasons. One is that in terms of your choice of access. Well, one simple answer is that you take all the internal data that are available. Essentially, this is throw everything by the kitchen sink approach. Often times you might want to augment your internal data with externally available data. So sometimes you may have to go beyond the boundaries of your organization to get data for an analytics project.

The next point is an important one. It essentially says that just because you have data does not mean that you

can or you should use the data. In many, many settings there are a host of legal, ethical and regulatory issues that will prevent you from using data. For example, in the U.S. when it comes to things like giving loans you cannot discriminate based on race, gender, ethnicity. If you have data about someone's gender, someone's race, their ethnicity you cannot use the data. That would become illegal. Now this question gets a little tricky because sometimes you cannot even use the location for example - zip code. Because in certain zip codes there are people with certain racial mixes tending to live more predominantly. So even though you're actually using location like pin code or zip code but in a sense, you're actually using ethnicity. That is also not acceptable. So you have to be very careful about the context of your application whether the access to predictors is ethically, legally okay. That's a question you need to ask. One other element I want to point out here is something called feature selection. A typical strategy is, you first throw everything applicable to the problem. Then as you look through it, you'll realize that some variables are important in prediction and some variables are not. So you begin to throw out ones that are not as important. So the objective here is to predict as well as possible with as few access as possible.

Just to give a quick example - Target in the U.S. figured out that based on data that they have they can actually predict which of their customers is pregnant and are in their first trimester of pregnancy. So what the data allowed them to learn is that apparently pregnant women in the first trimester have very unique purchasing patterns. For some reason they buy more cotton, for some reason, they buy perfumes and lotions which are unscented. So, they found that this very interesting pattern within the data and then that then changed their marketing strategy in terms of how they do outreach to this important segment of the market.

We've asked the why, when of data access. The next question is 'what' of the predictive model. Well at the core, the predictive model is a formula. In this example I have three predictors; gender, marital status and the age and my prediction is something called country.

Here the example is about cars. You want to predict whether this customer is going to buy a car, which is a Japanese made car, or whether they going to buy a car that is an American made car. You want to predict which of the two cars they buy. The predictive model here might look like this, simple one is if they're female I predict they go to buy a Japanese car if not they'll buy an American car. That's the prediction model. Here is another prediction model. If they are female they will buy a Japanese car. If they're male and if they are married they'll buy Japanese otherwise they'll buy an American car. At the end of the day, the prediction model is nothing but a formula. A formula that links your access, to predict your outcome Y which is of interest. Some prediction models are simple and some are more complicated. Now, this leads to the question would a more complex prediction model be better? Obviously, if you do something more sophisticated, more complicated that should give you a better outcome. In some sense that is true. Here's another example to illustrate this concept and in the field of analytics, there's an important concept called overfitting. This is an example that really illustrates it well.

In this graph, you see my X is just 1X of nitrate - the amount of nitrate content in the soil. And I want to predict how much yield I will get when you are sowing corn. A simple model is called a linear regression is a straight line model. As you can see here it predicts somewhat well but some things are really off.

Now I can make this model more complicated so that my fit is better. So maybe I go with, I go from a linear to a quadratic. It's flits slightly better with more of the data. The more complicated model seems to fit better. Now if I really want to fit this better I can make it a lot more complicated there's something called splines. As you can see here if I use a really complicated spline based model it fits the data really well. So as you can see from those three graphs at the top as you go from the left to the right the models get more complicated the fit gets better so I'm able to predict better.

The question is does that mean that you should always build as complicated a model as possible so that I have a really good model. Answer is, no. The reason is even though these complicated models fit well on existing data when actually applied they may not work that well. And the reason for that is the complicated models tend to fit all the idiosyncratic aspects of the existing data but they don't generalize them well. This an important concept in analytics called **'overfitting'**.

The way to overcome overfitting is that when you take data that you have access to, you split that into at least two groups. The first data set is used for building models and then the second data set is used for validation or for testing. Eventually, you build a variety of models and test them all on this fresh testing data to decide on whichever model does well.

Here's an example of this. There is a lot of data and this spline model has the least mode of error 10% error,0.01, whereas regression for the same data has 34% error. So obviously spline is a lot better. But when you apply this model to a fresh data set, you see that the complicated spline model does very poorly on the validation set. The error rate is almost 60% and if you compare the performance of all the models on the validation set, it looks like the quadratic model has the least amount of error. So that's the best model. In this case, what you see is that it's not the most complicated model that wins but something that's in between. That's the best. This is important concept.

In terms of how you build these algorithms, there are several different types of machine learning algorithms that have been developed and they continue being developed. But broadly speaking they fall into four broad categories. One is **regression** based models which is essentially a linear equation or sometimes, a quadratic equation. The second set of techniques are called **support vector** machines or support vector regressions. The idea here is when you have data which has some inherent non-linearity complexities in them, you can build better models by changing your dimensions. By altering the dimensions you can actually

begin to predict a little better. These are called support vector machines and they are very popular. And then you have **decision trees**. Decision trees are 'if-then-else' type of models. The most complicated models are called **neural network** models. These tend to be very very complicated and it's very hard to understand as well because there are a lot of non-linearities in them.

Now a few points I need to make here. One is that in some models you can write out your formula - your prediction modeling formula. Regressions and support vector machines can be written down like a formula. And then there are algorithms where you can follow it with and you can see the logic behind the algorithm. For example, a decision tree is a very logical prediction algorithm because you can take a look at it. Why am I saying this is a bad car? I'm saying this is a bad car because this happens, this happens, this happens then it's a bad car. It's easy to understand. Neural networks, on the other hand, are really hard to understand. This is highly mathematical. It's very difficult to really comprehend. So if your objective is comprehension, neural networks are not good techniques to use and in fact, in some application settings, you cannot use neural networks. Going back to the loan example, if your prediction algorithm is going to be used to either give loans or to deny loans you cannot use neural networks. Just imagine if a customer is denied a loan and comes back and asks why was the loan denied - if it cannot be explained, it is not good enough model to use. Neural networks work well in several settings but there are problems in explaining them as a process. The operate as a black box which prevents them from being used. The other problem with neural networks also is that it tends to overfit quite a bit. We talked about the case of overfitting before. It happens a lot with neural networks.

As we talked about before, we need to come back full circle to what are the business imperatives, why is the predictive model in the first place, what are you going to do with it, etc. So, the key here is to use the predictive model to make some important decisions for your organization and the idea there is to enhance the business value. Creating new business value is the most

important driver here. So how do you take a predictive model and then make a decision, optimize something so that you enhance business value? That's the critical question. And this is called prescriptive analytics. As I talked about before there are three broad elements to analytics; data management, predictive analytics, prescriptive analytics. So, the end result of your predictive analytics is the input to prescriptive analytic. So how do you make an optimum decision, how do you change your organization, what decisions do you change, what processes do you change to create value from analytics.

Here's an example and in fact the research that I will present will dig deeper into it. In an example of what a prescriptive analytics solution might look like in the context of Carvana, they try to predict kicks, bad buys. The end result might be something like this. If the probability predictive algorithm tells me that if the kick is between 0 and 0.4 very low likelihood of being a bad car then we buy it. From 0.4 to 0.6 there is a risk, but the risk is not too high but I still want to proceed but I want to bid a lower price. If it's between 0.6 and 0.7 perhaps I need to be a little bit more cautious so maybe before I make a decision, I should do some additional inspections in the car and if it's 0.7 and above it's too risky and I just won't be involved. So that's the outcome of your prescriptive analytics. Now optimization here deals with how do I come up with these cut-offs, how do I come up with the right choices and the objective would be to lower your losses due to kicks as much as possible.

Now one point I do want to make here is that in the field there are two communities there's what's called the machine learning community that focuses a lot on machine learning algorithms that we talked about before. There's another community it's called optimization decision analysis community that focuses a lot on optimization - how do we make optimal decisions for the organization. So what we're trying to do in my research is to combine machine learning with optimization. I think there are lots of interesting opportunities there. So that's what my research is going to be about.

Now before I get to my research let me just pause for a second. I'm going to stop the sharing. See are there any questions. I'll answer the questions and then we move on.

One question is '*how can we eliminate biasness when throwing out some variables?'* This means, how can we eliminate biasness while throwing out some variables?

When you throw variables, there are a number of criteria that you have to keep in mind. Obviously one of the criteria is bias. You have to make sure that by throwing out some important variables you don't enhance bias. In terms of feature selection bias, deduction is one of the objectives. There are a couple of objectives in terms of feature selection - one is to improve performance. Second is to lower the degree of bias. So that has to be part of the equation in terms of the choices you make in terms of what variables you throw out.

The next question is '*what's the difference between validation and test data sets?'* Again, I did not want to get too much into the technical details but the difference is that with validation the idea is to pick the best model. We have a set of competing models that you throw into the problem and as I mentioned before, the objective is to pick the best model and in terms of picking the best model validation test data set is used. The next thing is after you pick the best model you still have to fine-tune. For example with the kick example that I told you, there are different cut-offs. The question is how do you make those optimal cut-offs and for that, we use the test data set.

So, both of them are important. They follow the training data set but with validation the objective is to pick the best model. The testing data set the objective is to fine-tune the best model that you have picked.

The next question here is, '*when Amazon is putting a price to a prospective buyer, is it using prescriptive analytics or predictive analytics?'*. Again I'm not exactly sure what the setting here is in terms of Amazon on

putting a price on a prospective buyer. Now again the objective is that taking a look at individuals and if you want to see that a prospective buyer or not then the approach obviously is predictive analytics and that's what you would look for. On the other hand, if you have data on the whole host of customers and then your objective is that what group should I target, should I be targeting males under this age or females in that age range things like that then you would use more unsupervised learning.

'Do we use any tools to measure cut offs in prescriptive analytics?' Yes. I mean that's where optimization comes into play. The important thing there is to look at the cost and benefits. What are calls the cost of type 1 and type 2 errors? Again in the example of a car being a kick or not, what if you make a wrong decision? It's actually a good car but you're not buying it or it's a bad car you make and you actually buy it. The different types of cost that you would incur tend to optimize and then come up with the right cutoff.

Next question here is what are the parameters we should keep in mind while selecting variables in the prediction model? Well variable selection and feature engineering, feature selection is an important one and as I mentioned the most common strategy is to start with whatever you have access to and you begin to pare it down to have as good a performance as possible while throwing out variables that don't add much value. You should also mention that the number of other techniques as well because sometimes instead of throwing things out what you might do is to combine all the variables into one something called principle components and factor analysis the techniques like that to again try to reduce the number of variables that you use.

The next question here is *'how to separate noise and unwanted data from a data set which is messy and non-uniform?'* I think you're talking about data cleansing. That's again we did not get to talk about it but that actually is a very important step in predictive analytics. The data that you often have tends to be not very clean and there are the whole host of issues. There could be

errors in the data. There could be missing data and data coming from different sources might have different meanings. So, it's not apples to apples. There's a whole host of approaches that fall under data cleansing that is used to address these questions.

One question is '*there are a lot of external factors and influencers that are not captured in the data and not being aware of these factors might throw us on the wrong path. So how to tackle this?*' That's a great question. You don't know what you don't know and, you don't know what you know essentially. If there are some important factors or data that you simply don't have access to, the only way you can learn about them is by looking at your predictive model. If your model is missing something really critical or important your predictive model will not be that great. That will be your first clue as to you that you might be missing something that is really important. There are several phenomena like '*why does someone not repay the loan?'*. There are some clues you can get from the data you have about the customer but there could be so many things happening in that person's life that you don't know. That could also be an important driver. So, these kinds of external things have an effect. But from a business point of view, the question is that '*can you predict well enough that you can do something about it*?'

****