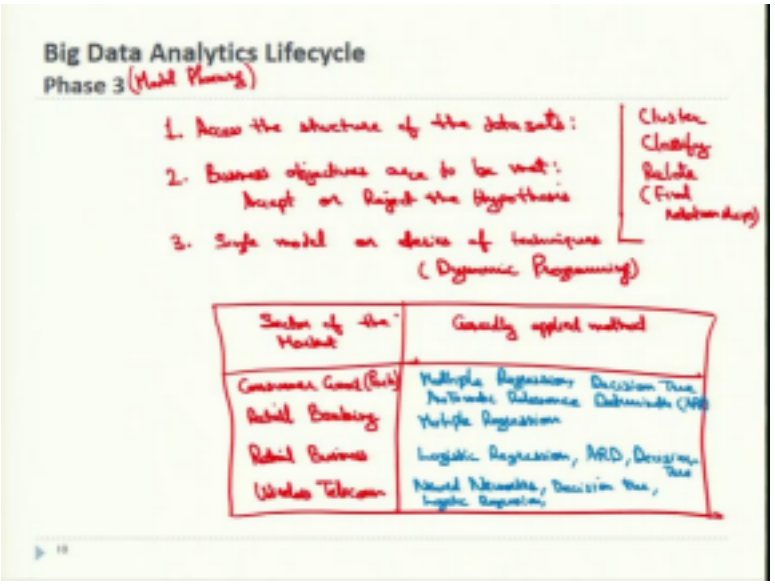
Computer Aided Decision Systems
Industrial Practices Using Big Analytics
Professor Deepu Philip
Department of Industrial and Management Engineering
Professor Amandeep Singh
Imagineering Laboratory
Indian Institute of Technology, Kanpur
Lecture – 21
Big Data Analytics Lifecycle-Phases 3 & 4

Welcome back to the course on Computer Aided Decision Systems, where we are trying to learn Big Data Analytics, and I am discussing the Lifecycle of the Big Data Analytics System. We have discussed two phases. In phase 1, we have seen how we discover the data? Who, what, how do we understand the Business Domain? In phase 2, we had a small introduction on how to prepare the data.

For Data Preparation, we have to interview or we have to meet the sponsor. We have to conduct ETLT analysis, and we also prepared the Analytic Sandbox.

(Refer Slide Time: 00:50)



Phase 3: Model Planning

It is still planning the model, we are not building the model. In the Data Preparation, as it was said, we have invested the maximum time to prepare the data, to collect the data, and to put that in the form, to put it in the right transformation form that could be used in the further phases, phase three or four, we will try to build and we will try to plan the model. But still,

Before building the model we have to train the model, we have to see how we plan or how we build the action model, which is a model to communicate it further to operationalize it later.

So, Model Planning is the phase that would take the input from phase two, and here the major things that we do are the clustering, the classification, the finding relationships between the datasets, and we try to also refer to the hypothesis, which were developed in phase one.

So majorly, we try to,

1) Access the structure of the datasets- When I say structure of the datasets, whether the data is structured, unstructured, is it available in a formal form, means it is taken by some software or so? Or is it available in just the unstructured text, which has to be converted into formal text? And what are the opinions of the reviewer or the person who was given the review? So, we need to access the datasets. The structure of the datasets is one of the factors which determines the tools and analytic techniques. In the next phase that is in the Model Building phase.

So, it depends whether the team plans to analyze the textual data only or the transactional data. For example, different tools and approaches are required for different kinds of the datasets which are available. So, here we majorly cluster, classify, and relate the data. When I say relate, we find the relationships. Now in this phase, the team references the hypothesis which is developed in phase one. This Hypothesis helps a team to formulate the analytics, which is to be executed in the Model Building finally.

So, when the accessing of the data is already taken care of, then we make it certain that the analytical techniques are enabled so that the team meets the business objectives and accept the, or reject the Hypothesis.

2) Business objectives are to be met- We plan whether the Hypothesis will be accepted or it will be rejected. We can test the Hypothesis such as, investment in the innovation research in the company, the increase in investment made in that duration by five percent would lead to the ten products developed in a year, that is all. This is the Hypothesis that is tested.

We try to accept or reject this number that is put here, ten would be developed, five percent

investment. I am trying to put very specific numbers here. These numbers could be a little varied if required. If you see that in the training data itself, if it is not meeting or it is not, the Hypothesis is not being accepted. So, then we try to change the Hypothesis statement even to finally build the model. So, this is where we majorly, we are trying to train the data.

So, this Hypothesis could be different for different companies. For the retail companies like Walmart, the Hypothesis would be, all the sales would increase, or the customers traffic is to be controlled, or the time to, the customer spends in the queue is to be reduced to the minimum. Or how do we stack the things so that, everything, maximum display is there for the items which are there. And how do we deliver them in time, in a timely manner? Like people who come to the store itself or people, and then in the Covid's time, they had spare baskets outside the store only. So, the order was taken online and in the baskets itself, they used to put the system or the, whatever the order was, so the customer used to come and pick it from there. Now itself, in Canada, I have observed that people still take the orders from the baskets themselves outside the store. So, the person from the store comes outside and delivers it there.

So, what is the delivery time for that? What is the waiting time of the person in the parking lot? So, all these could be put as a Hypothesis for different companies of different types of objectives that we are trying to meet. Business objective is, how would this value to the customer? How would this increase the customer's loyalty? Loyalty is a very broad term. That is also to be brought into a quantitative manner. Maybe loyalty is increased from the rating, from a loyalty rating from 1 to 10 percent, maybe 4 percent. Now, he or she is giving the rating seven or eight. So, how is the loyalty of the person increasing in a way? So, these things could be put or could be changed.

3) Single model or a series of techniques- This means whether the program is just a deterministic model or maybe a small probabilistic model that we are trying to develop or is it a completely dynamic model, whether things will keep on changing as and when the model progresses? So, here it becomes a series of techniques. It might be Dynamic Programming as well. So Dynamic Programming in itself becomes a little complex, it is only the extreme case when the system is too versatile. Versatile in a way when the velocity of the data is a variety of the data and the velocity of the data is heavy, so the data is also unstructured. So, really Dynamic Program could also be

taken care of in the Big data programs.

So, these are the major concerns which are there in phase three. So, in addition to these considerations which are listed here, that is 1, 2, 3, it is useful to research and understand how the analysts generally approach a specific kind of problem? So, for example, if it is given a kind of data and research which are available, then we have to evaluate whether the similar existing approaches which have been applied into the similar processes, would work in our system or not? Is the behavior of the local public similar, is the environment within my concern or within my company similar? So, things would change.

For example, if some program is successful in the US, and in India, because the people who are working here might have different perceptions of it and customers also have different paying capacity or spending capacity. So, the programs might be a little different. So, it could be a little different, it could be largely different. So, we have to see how we change that. So, This is also determined in the Model Planning only.

So, we get the ideas from the analogues problems that the other people have solved in different industries or different verticals and we try to summarize that.

Majorly, if I jot down what kind of sectors and what kind of techniques you know, in statistics there are certain techniques like Multiple Linear Regression we call it. Then Logistic Regression, we have cluster analysis, we have correlation analysis only where the causality is not taken. There are certain techniques there.

So, generally if I put a small table here, I would say, sector of the market and generally applied method. So, I will take these sectors here. So let me say, if it is consumer goods, like we took in examples of Walmart or the similar stores there. Our second sector is retail banking. Then maybe retail businesses. Retail business is actually the store business. So, it is consumer goods, we packaged, number 1 is packaged. Then wireless telecom. So here, in the case of the consumer-packaged goods, the program that would be majorly used would be Multiple Regression only. When I say a word Regression, that in itself means we are only using the Linear Regression. Polynomial Regression is not yet used because the pro model becomes a little complex or there is laziness, there is a repetition.

So, the higher order aggression is not there. It is multiple digression; Multi-objective

Regression could also be there. So, objective could be either one, or objective could be more than one. If more than one objective is there, we could even use some heuristics. Like we can use Genetic Algorithm, or Artificial Bee Colony, relational analysis that could still come. But Multiple Regression is the base technical method that is used. So, other than that, the decision tree cannot be missed. Yes, Automatic Relevance Determination (ARD) techniques are also there. It is ARD.

In the case of retail banking, because we are straightforward with the things we were coming up with. So majorly, Multiple Regression only works. In the retail business, majorly what we work upon is Logistic Regression. Then ARD techniques and decision tree.

How is Logistic Regression different from the Basic Regression? In Basic Regression, the variables can take any value. It could be, it is okay in the Logistic it so all can be value, but the predicted value that is taken in the Linear Regression is generally mean of the target, or it is generally some average or so. In Logistic Regression, we generally classify, yes or no. The email that has come to your office is spam or not spam. The credit card transaction is fraudulent or not fraudulent. So, this is Logistic Regression, where we want to try to only classify the things. So that is why Logistic Regression, the name is called it.

So, ARD, relevance techniques are the one where we try to employ the Bayesian techniques. That is the Bayesian interpolation or determination, the probability based upon this kind of redistribution, Bay's distribution only. So, decision three is one that you have learned in the previous lectures. Doctor Deepu Philip has given you a brief introduction of that.

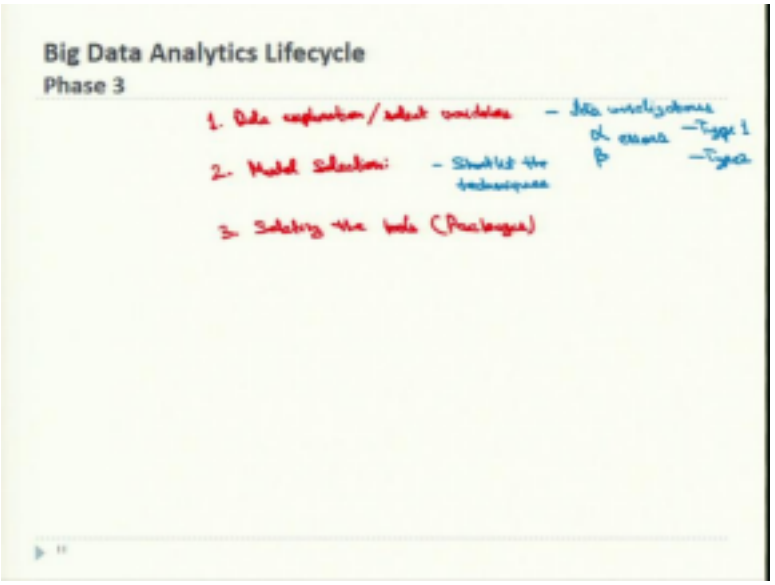
In wireless telecom, there are multiple things that can be taken care of because wireless telecoms do have the data. They have different connections, the people who are connecting, the time for the call. Then the rate for which the call is given is different mobile services, which are there. So, here definitely are the most sophisticated techniques like we have Neural Networks, Decision Trees. We have to also classify whether the call is even fraudulent or not, or spam or not spam. So, then we have to write here again, Logistic Regression. So when, if I have put a Neural Network, that means other algorithms such as Artificial Bee Colony, the Genetic Algorithm, those could also come here. These are the major classifications that I put, generally which are found more applicable in these kinds of the applications, it is kind of the sectors in the market. But yes, it is interchangeable, if the problem statement is in such a way

that even in the retail business, we are not only talking about the spam or not spam or fraud or not fraud, we are even talking about what is the percentage increase or what is the number increase.

There also, the Linear Regression or the decision tree or the ARD techniques could also play. So, it depends upon what kind of problem statement, how are you defining the hypothesis. So, it can be interchangeable. So, the gist of the talk or gist of this table is that there are major techniques which are generally applied, which are approved from the past applications of the profit examples. But you can only understand, what are the major techniques, what to use where, once you are aware of the techniques and you are versed or well versed with the different methods.

How do we use them? It will itself come. So go, softwares or the systems are always Garbage In Garbage Out. So, a Business Intelligence Analyst has to understand what kind of different techniques are available? What are the media tools which could be used to those, could be applied? I will talk about the tools and techniques in the forthcoming slides as well. So, if we have jotted down the steps in phase three, what is to be done in phase 3, that is listed in the previous slide.

(Refer Slide Time: 16:21)



The steps in phase three would be,

1) Data exploration/select variable- This means the objective of the data exploration, that is, to learn the relationships among the variables so that we understand the problem

domain properly. So, a commonly used way to conduct this step is involving tools to perform data visualizations. Approaching the data exploration in this way aids the team in previewing the data, assessing the relationship between the variables at high level.

So broadly, we can put small visualizations here, so that we can see which of the variables are relevant and which are not yet. So, in most of the cases, these stakeholders have instincts or hunches about what the data science team should consider analyzing.

So, often stakeholders have a good grasp of the problem and domain, although they may not be aware of the, maybe small nuances of the data, or the Model needed to accept or reject the Hypothesis. The second case could be that the stakeholders may be correct, but for the wrong reasons. So, the correlations have to be made. We call it the Type 1 or Type 2 arrival. Type 1 error is the α error. That is, the variables which should be there, which are going to represent the model in the realistic form are rejected, because the analytics or the tests that you have conducted were not correct or you chose the test, which were not correct. This can be a Type 1 error.

Type 2 error is that the test you have conducted on the sample dataset that you have taken, so you have selected the variables, but these variables are not representing the final model or you are not representing the realistic form of the final model. So, this α or β error could both come. Which means sometimes they overdo the things, sometimes they undergo the things. So, both are not allowed.

So, data exploration and selecting variables, both of these errors are to be taken care of. So that is why we draw some correlations that are in data visualization. Majorly we draw the correlations and try to see what is the correlation, but what is the reason why the correlation, the causal relations since do not come, for that, we have to take further analytics.

Now here the team begins to question the incoming assumptions of the initial ideas, initial Hypothesis. The assumptions might not be correct at which, might be one of the reasons or it may be, it might need to consider then alternate matter or reduce the number of data inputs or transform the inputs into, into the way so that the team best understands it, so best understand the business problem. So maybe the basic business problem is being missed. The team is diverging from the basic business problem.

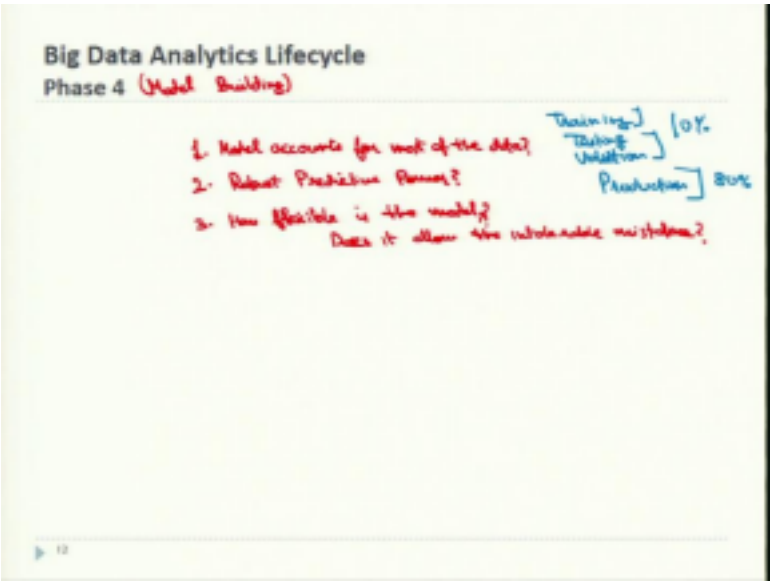
If it is a 5 percent investment in the innovation strategies. So, the five percent investment might only be made in purchasing the equipment of the innovation, of the five percent investment might only be made in hiring the people only, but actually if it is a five percent investment, it has to be distributed properly. The business remains ten times increase in the products or the innovative products per year. So, for that, what is the right input, the investment is made in which sector of the company? So that is also one of the points that for that data points also might have to be collected. So, it depends, how we understand total business, overall business domain, and overall business results. So, this is how we go about it.

So, the key to this approach is to aim for capturing the most essential predictions and variables rather than considering every possible variable and people may think that might have influenced the outcome. So, it requires a lot of alterations and testing to identify and the team should plan a range of variables to include the model and then focus on the most important error on your influx.

2) Model selection- Model selection means here the analytical techniques or shortlist of the candidate techniques are there. Then we try to set the rules that, if one observes events happening in a real-world situation or with live data, attempts to construct a model that would emulate this behavior with a set of rules and conditions. So, in the case of machine learning or data mining for example, the rules and conditions are grouped into several, general small sets of techniques such as classification, association, rules, clustering, so that we try to solve the given problem. So, the data could be structured data, unstructured data, hybrid approach, we try to identify and document the model's assumptions here. The initial models are now taken through some software such as R, SAS or MATLAB.

3) Model Building- Here we are selecting the common tools. When I say tools, I am talking about the software packages. This I will talk about when I complete all the phases.

(Refer Slide Time: 22:20)



Phase 4: Model Building

In the Model Building phase, we develop the datasets for training, testing and production purposes and we try to develop the analytical model so that the training data will try to train it. That is, it will try to see whether the initial tests are okay to clear.

Then holding aside some of the data that we call test data for the testing model, then we try to validate that and we have to go for production. It is not only training, testing, when I say testing, it includes validation. And then we go for final production. Here, when I am talking about training and testing, the major data might be 80 percent, 90 percent of data is taken for the major production, but this is only 10 percent of the data that is used for training and testing only.

So, in general, a plan to spend more time preparing and learning the data that was in phase 1 and 2, is imperative and it crafts a presentation of findings. So, phase 3 and 4, moving quickly to them is not helpful. So, it becomes more complex unless we have prepared the data and we have planned for the model. So, it is critical to ensure that the training and test datasets are sufficiently robust for the model and analytical techniques.

So, we need to determine two things here majorly.

1) Whether the Model accounts for most of the data?
2) Does it have a robust predictive power?
3) How flexible is the Model?

For instance, if I am picking the Regression Model, learning Linear Regression only, the output would be a quantitative variable, and I need inputs, which would also be quantitative. I need metric data, but my data which is available is only in the textual form, which I have not converted into the quantitative form or data is only available in the qualitative form, yes or no, or so that.

So, does it account for the most of the data? So, what is the maximum amount of data that we have available? Do we need to transform that data into usable form or should I select something X? Should I select maybe a classification model only? Should I select maybe a Cascade test? Or should I only select a Logistics Regression, which could be more applied on yes or no Models? The Predictive power means the prediction that has come, would that also represent the final results to be representative or to be communicated to enough to serve that phase or not? These two are the major concerns in this phase. So, one must take care of the record or any operating assumptions that were maybe in the previous phases. So, this is how it goes. So, we create a robust model.

The second question is covering robust Predictive power, it means that the parameter values, which are put in the model, make realistic sense in the consent of a domain, in the Business Domain that we are selecting. So, is the model sufficiently accurate to meet the goal? So, does the model avoid intolerable mistakes?

The third question says that how flexible is the model? That means, does it allow intolerable mistakes? So, then if we suppose it needs to change something in it, or more data, more inputs required. Is the model able to accept more data while running itself? So, can we have a parallel running between the model and data ingesting itself? Is a different form of model required to address the business problem? All these questions could also be asked. So, this is a quick view of phase 4. Thank you.