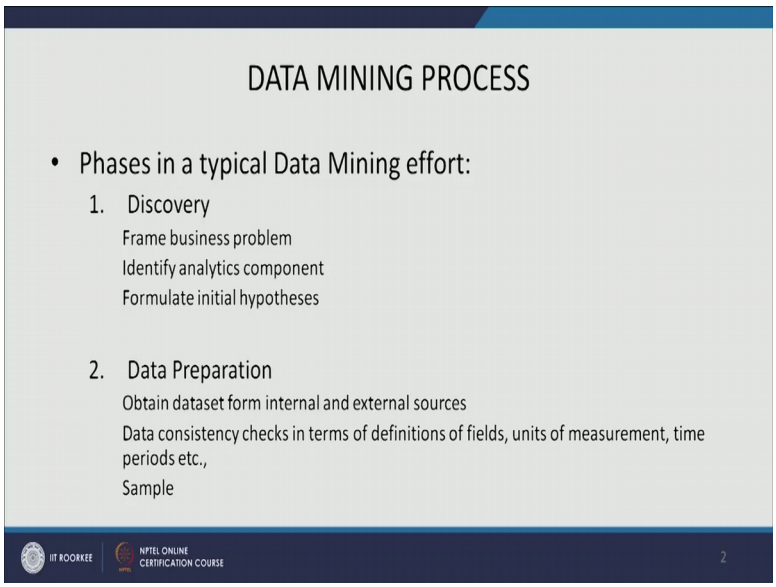
Business Analytics & Data Mining Modeling Using R

Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture - 02
Data Mining Process

Welcome to the lecture 2 of course Business Analytics and Data Mining Modeling Using R. In this particular course, in this particular lecture we are going to cover the data mining process.

(Refer Slide Time: 00:30)



So, there are different phases in a typical data mining effort. So, first phase, phase is about first phase is named as discovery. So, important activities that are part of this phase are framing business problem. So, first we need to understand the needs of the organization and the issues they are you know facing and the kind of resources that are available the kind of team is available, the kind of data repository other resources are available, and the you know try to understand the main business problem and then try to identify the analytics component, so analytics challenge or analytics components, component part of that particular business problem. Then we start understanding we start to develop our understanding about the relevant phenomena the relevant concepts constructs variables and we try to develop or formulate our initial hypothesis that are going to be later on converted into a data mining problem.

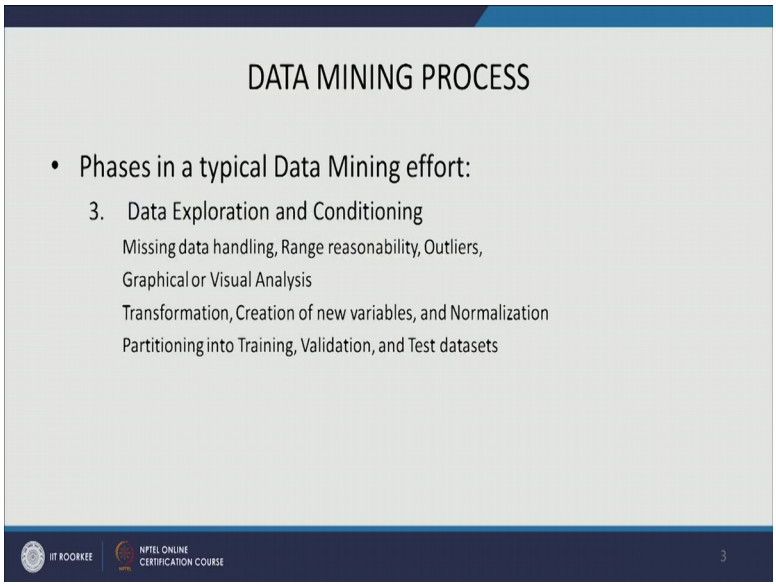
So, this is in the discovery in the first phase these are some of the activities that we generally have to do. The next stages, the next phase is a data preparation in this phase we have already understood the problem especially the analytics problem that we have to deal with therefore, we also based on our initial hypothesis formulation we will also have you know understanding about the kind of variables that would be required to perform the analysis therefore, we can also we can always look for the relevant data from internal and external resources and then we can compile the data set and that can be used for the analytics.

So, another activities that can actually be performing this stage also a data consistency checks. For example data can come from variety of sources therefore, we need to check the definition of fields whether they are consistent or not we need to we also need to look for units of measurement we also need to check for you know data format whether that is consistent or not. For example, if you have a variable gender and your data set and in one particular source is it is recorded as male m a l e and in other male and female full form m a l e and female and in another from another source it is recorded as capital M or capital F. So, therefore, you have to make sure that data is consistent and when the whole data is compiled or got together and, these consistency have to be checked.

Then time periods. So, data could be belonging to a particular you know you know number of years so that we also need to check because the analysis or results could actually be limited by the time period as well. Sample, now you do not always need all the records that we prepare in our data set we normally generally we take a sample of it because smaller sample of the you know that is generally good enough to build accurate models.

Now a next phase in a typical data mining effort is about data exploration and conditioning.

(Refer Slide Time: 04:05)



So, in this phase we generally do activities like missing data handling we also check for range region ability whether the range of different variables they are in the expected, they are as per the expectation or not. We also look for the outliers they can come due to some human error or measurement error etcetera. Graphical or visual analysis is another activity. So, we generally you know plot a number of a number of graphics for example, histogram, scatter plot, that we are going to discuss in later coming lecture other activities that we can do is transformation of variables, creation of new variables and normalization.

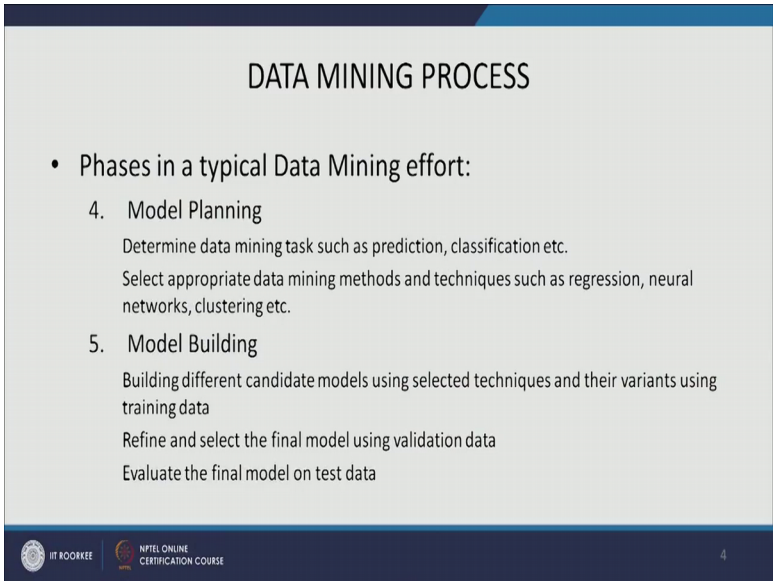
So, why we require, why we might need to do some of these activities would be more clear on as we go further in coming lectures. Transformation creation of new variable for example, sometimes you know we might require that you know sales figures has to be you know in more than 10 million or less than 10 million kind of format then we might have to transform our variable. Creation of new variables, if we do some of these transformation we will end up with some of the new variables sometimes we might use both the forms in our analysis. Normalization sometimes a scale could pose a problem in a particular data mining algorithm or statistic call technique, therefore, normalization might be required. So, these are some of the activities that we have to do.

Partitioning, partitioning is another essential part of a data mining modeling. So, be generally partition our data set our sample into training were relation and test data sets.

Training is generally the a large partition and all the models all the elevation they are generally applied on the training data set and then fine tuning or selection of a you know particular algorithm or particular technique happens on the validation data set test data set is again reevaluation of that final method.

Next phase of data mining process is model planning.

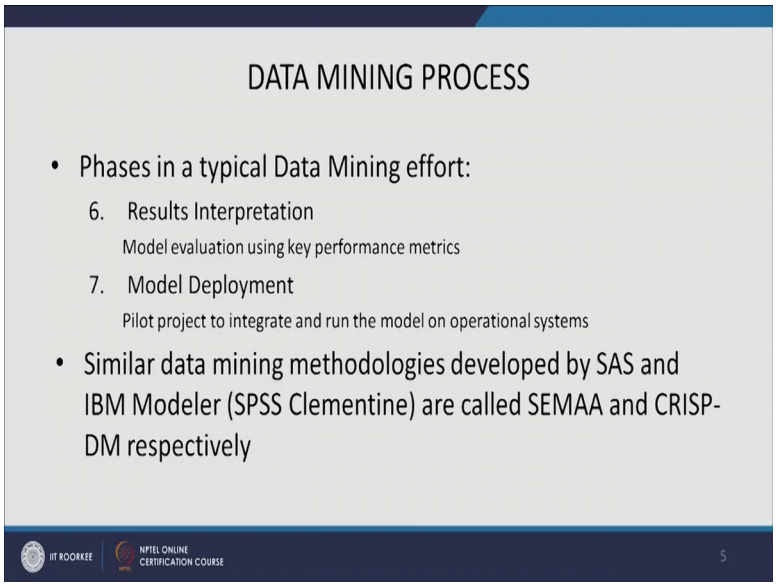
(Refer Slide Time: 06:27)



So, in this particular phase you know we need to determine the task that we have to perform whether it is a prediction task or a classification task. We also need to select the appropriate method for that particular task if it could be regression neural network, it could be clustering discriminant analysis and many other methods that we are going to discuss incoming lectures.

Now next phase is a model building, building different you know candidate models or using or selected techniques in the previous steps and their variants. So, that is run using training data then only we define and select the final model using the validation data then evaluation component that happens the final model that is done on test data. So, in this phase mainly it try out different candidate models we assess their performance we find tuning them and then we finally, select a particular model.

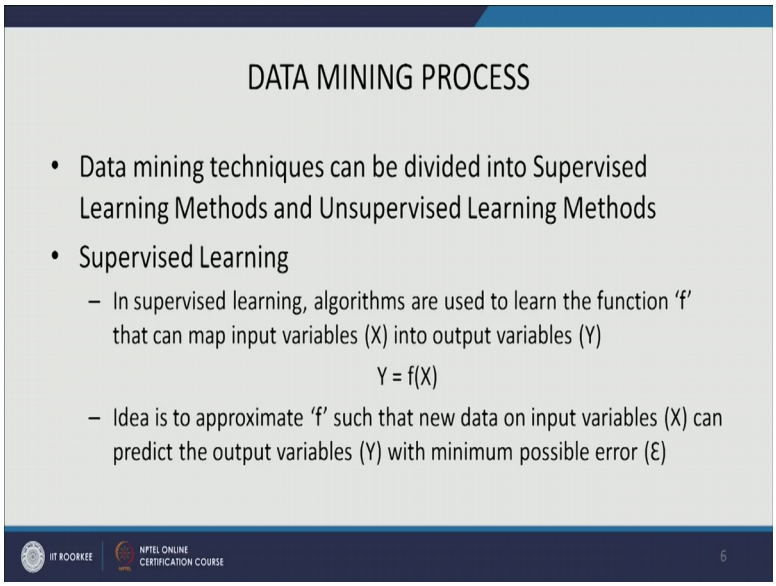
(Refer Slide Time: 07:24)



Now, the next phase is results interpretation. So, once model evaluation you know happens using performance matrix we can go on and interpret the results. So, one is the exploratory part then another one could be the prediction part. So, all those interpretation can actually be done in this particular phase, and they then the model deployment once we are satisfied with the model that we have we can run a pilot project to. So, that we are able to integrate and then the model on operational systems. So, that is the last phase.

Now, this is the typical data mining process that one has to follow. Similar data mining methodologies were developed by SAS and IBM modeler that was previously known as spaces (Refer Time: 08:10). So, these are the commercial software for a statistical modeling and data mining modeling and they also follow a similar kind of data mining mythology. So, for SAS methodology is call SEMAA, S E M A A and the IBM modelers methodology is called crisp DM.

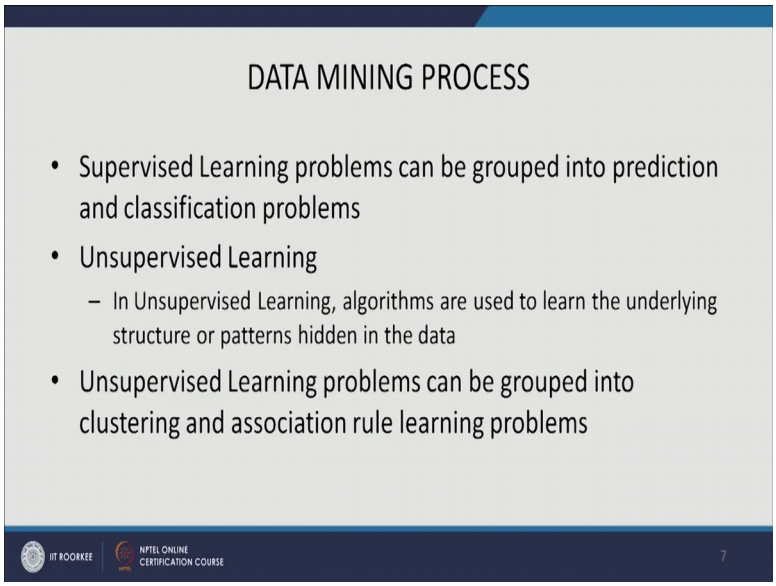
(Refer Slide Time: 08:29)



Now, at this point we need to understand the classification, important classification and related to data mining techniques. So, data mining techniques can typically be divided into supervised learning methods and unsupervised learning methods.

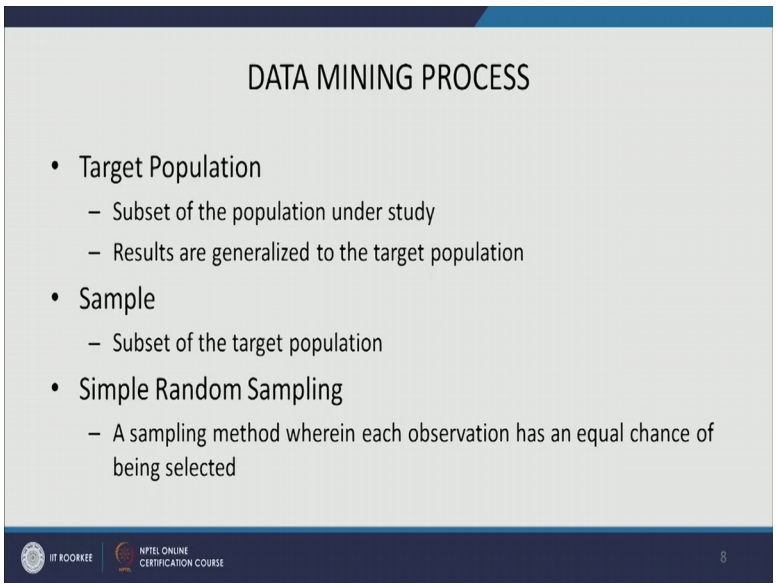
Now, what is supervised learning? So, in supervised learning algorithms are used to learn the function f that can map input variables that is generally denoted by X into output variables that is generally denoted by Y. So, algorithms they are used to learn a mapping function, a function which can actually map input variables X into output variables Y as you know you can also write this as Y as a function of X. Now, the main idea behind this is we want to approximate of mapping function f such that new data on input variables can actually be used to predict the output variables Y with minimum possible error. So, that is the main idea. So, the whole model development that we do is and this you know learning of this function is actually perform. So, that on new data we are able to predict the outcome variables with minimum possible error.

(Refer Slide Time: 09:48)



Now, supervised learning problems can be further grouped into prediction and classification problems that we have discussed before. Now, next is unsupervised learning. So, in unsupervised learning algorithms are used to learn the underlying structure or patterns hidden in the data. If you look at the examples again unsupervised learning problems can be grouped into clustering and association rule learning problems.

(Refer Slide Time: 10:17)

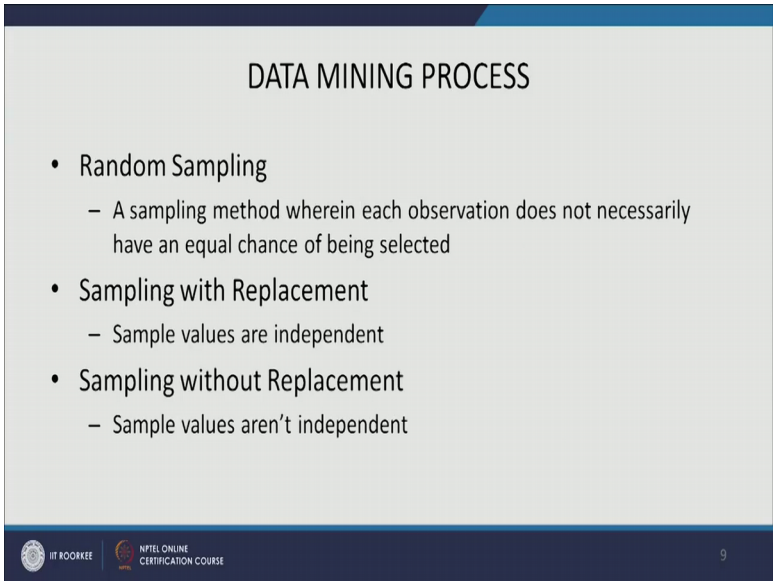


Now, there are some other key concepts that can actually be we need to discuss before we move further.

So, some of these concepts are related to sampling. So, we do not need to go into detail of you know all the sampling related concepts, we will need what is mainly applicable what is mainly relevant to a data mining process. First we need to understand the target population. So, how we define a target population? Target population is the subset of the population under study for example, in our sedan car example in the previous lecture it is the household that was the target population. So, we wanted to study the household and whether the own a sedan car or not. Now, results whenever we study a particular target population it is generally understood that results are going to be generalized to the same target population. Now, when we do an analysis we do not gather all the data of all the data coming from the target population we take a sample of it, reason are as we have discussed or indicated before cost related problems we cannot actually go about collecting data from the whole population because it is going to be a very costly process.

So, therefore, the purpose of you know business analytics data mining modeling and other related discipline is to reduce this cost and still have some useful insights or solved some of the analytics problem. So, that is why we require sample. So, we generally take a subset of a target population and then analyze the data that we have in that sample. So, generally our data mining scope generally we might be mainly limited to simple random sampling which is a sampling method where in each of observation has an equal chance of being selected.

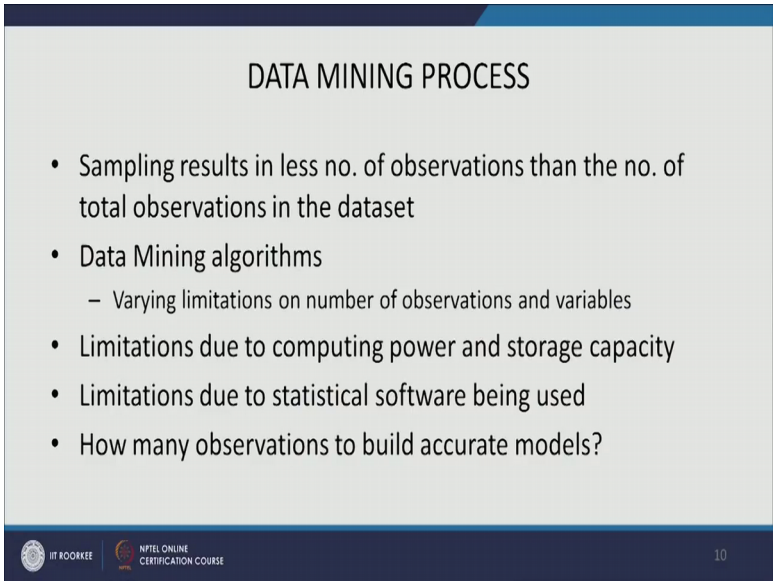
(Refer Slide Time: 12:27)



Now, what is random sampling? So, an sampling method where in each observation does not necessarily have an equal chance of being selected. Generally simple random sampling is used where each observation has a equal chance of being selected. The problem with simple random sampling could be that sometimes the sampled population could not be there represented of, could not be the proper you know representation of the target population because of this you know equal probability of you know equal probability of each observation being selected this could be one problem.

Sampling, next is sampling with replacement. So, when we do sampling with replacement. So, when we pick an observation then that observation is placed back in the sample then again for the next value it again has the equal chance of being selected. So, in sampling with replacement, sampling values are independent. So, once an observation is has means you know once and observation or case has been selected again the next observation does not depend on this the first selection. It can again be the same case or some other case. So, sample values are independent. When we say sampling without replacement sample values are independent because once a particular case or observation has been selected because of this without replacement procedure we cannot put it back there. So, therefore, the remaining observation we have to select from the remaining cases therefore, the next selections are going to be dependent on the previous selection. So, that is why sample values are not independent.

(Refer Slide Time: 14:24)

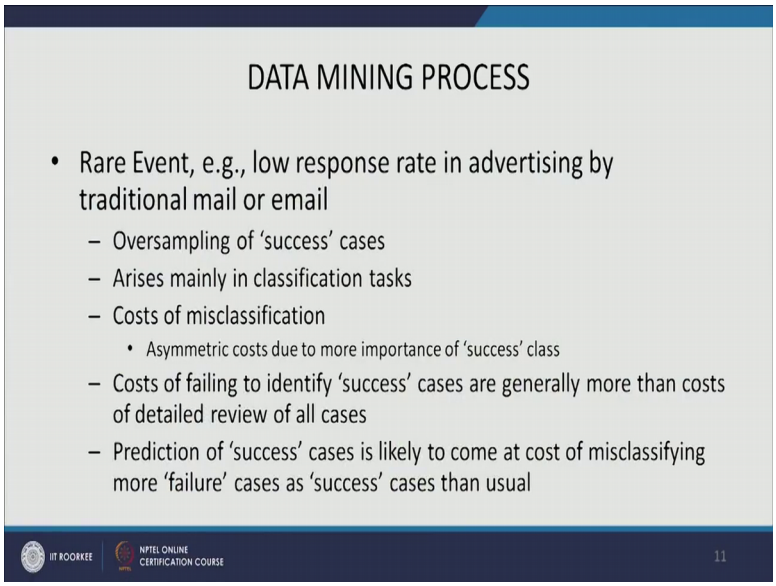


Now, whenever we do sampling it is going to result in less number of observation than the total number of observation that are present in the data set. Now, there are some other, there are some issues that could again further bring down the number of observation or variables in your sample. For example, data mining algorithm, so different data mining algorithm they could have varying limitation on number of observation on variables. They might not be able to handle you know more than a certain number of observation or more than a certain number of variables so that could be one limitation that could again limit your sample size.

Now, limitation can also be your due to computing power and storage capacity. So, the available computing power and storage that you have for your analysis purpose, can also either lower the speed or limit the number of observation the or number of variables that can be handled. Similarly elimination can also be due to a statistical a software, nowadays you already understand that different software they have different versions. So, generally if you are using a free version of you know commercial software that can actually limit the number of variables or number of observations that can actually be studied. So, those limitations can further bring down your sample size.

Now, while we are discussing this about you know limitation related to number of observations we need to understand how many observations are actually required to will accurate models because the whole idea is to build accurate models build good models. So, that we have good enough results which could be used on production systems later on. Especially when we understand that the, you know we cannot get the data from the you know target population we have to take a sample. So, cost is an important factor. So, therefore, it is always you know better for us to understand the number of observation number of you know sample size that would be sufficient for us to build accurate model robust models.

(Refer Slide Time: 16:32)



Now, there are many other concepts related to you know data mining process that we need to understand at this point, sampling size we will come back again. Now, next concept is rare event. Now, typically when you are dealing with a particular data set now if it is you know classification problem and you have a outcome variable where you have you know for example, ownership that we talked about whether a household owns a car or not. So, owner you know owner or non owner. So, that kind of scenario is there.

So, typically you know you know the split between observation belonging to owner class and observation belonging to the non owner class there would 60 40 or around 50 50 or you know that kind of ratio. But it might so happen that in a you know in a if it is rare event that out of let us say 1000 households only 10 or 20 household actually own a sedan car, then in that case it will make it a rare event, that ownership of a sedan car could actually be a rare event in a particular target population. So, in that case how do we do our modeling.

Another example would be low response rate in advertising by traditional mail or email. So, you might be having different promotional or marketing advertising offers through coming to you through traditional mails post and emails as well. Now not everyone is going to respond to this offers. So, again this can also be a rare event. So, how do we do our modeling? What are the issues that we face in this kind of situation? So, if you have a particular class which is which has very few observation belonging to you know very

few observation actually belong to this class. So, any kind of modeling that you are actually going to you know do using that particular data set might not give you an you know good enough model right, it might become very difficult for you to get a model build a model that will satisfy your main analytics goal.

For example, if you have 100 observation in your data set and 95 belong to the non owner class and 5 belong to the owner class and main objective of your business problem is to identify people who are owners. So, therefore, if you even if you do not build a model a still you classify every household as a non owner still you will get 95 percent accuracy of your model. So, in this case modeling for the success cases is becomes an issue. So, how do we solve these problem we do over sampling. So, we do over sampling of success cases so that means, we include more we duplicate many of the data points which are belong to the success case or we change the ratio that means we take the success cases data points observation and we also take the observation which belong to the for example, in we take observation belonging to the owner class and we take the observation belonging to the non owner class in a 50 50 kind of ratio or similar ratio.

This kind of problem again arises mainly in classification task now in other related concept to this problem is cost of this classification. So, when we say that success class is more important for us we are dealing with asymmetric cost here and because identifying success class is more important for us because it is more important for us to understand which customer is going to respond to our email you know promotional offer and marketing offer. So, therefore, we dealing essentially dealing with asymmetric cost.

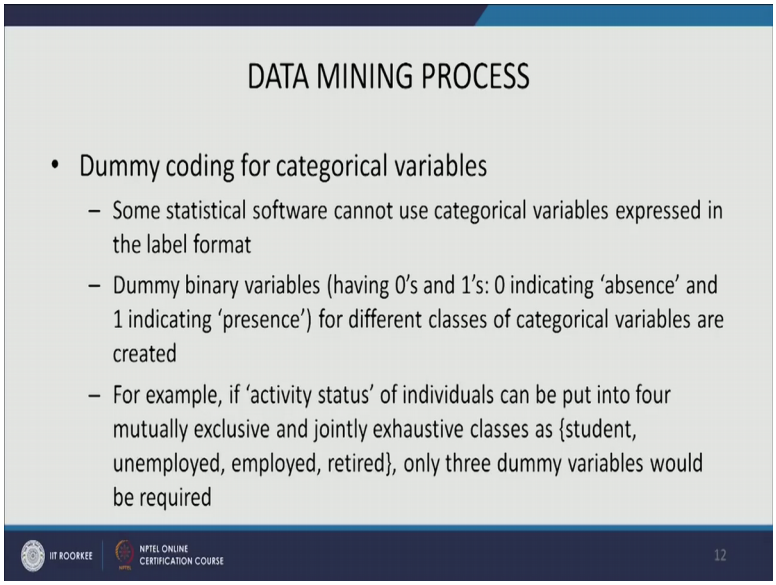
Now, generally if we are not able to identify success cases generally the cost of failing to identify success cases or is going to be more than cost of detailed review of all cases region being, if you are able to identify a particular customer that he or she is going to respond to your offer then probably they are profit that you can earn by selling the product or services to that customer would actually be more than cost of detailed review of cases.

So, therefore, generally the benefit of identifying success cases is higher than that and that is why modeling is modeling becomes premium. Now, other important aspect now other important aspect of rare event is when you know success case is important for you now prediction of success cases is always going to come at a cost of misclassifying a

failure case as a success case. Because if you out of 100 observation you have only 5 success cases you would like to identity all those 5 so that you are able to make a profit from your offerings. So, therefore, eventually the model that you built you might end up with you know identifying many failure cases also as a success cases. So, therefore, this is going to be more than usual so, is usually if you have you know 50 50 case of success case and failure success cases and failure cases your accuracy or your accuracy and error can be on the error can be on the lower side, but in this case error would actually increase, but the purpose is to identify success cases to in increase the profit.

Now, another important aspect of data mining process is dummy coding for categorical variables.

(Refer Slide Time: 22:35)



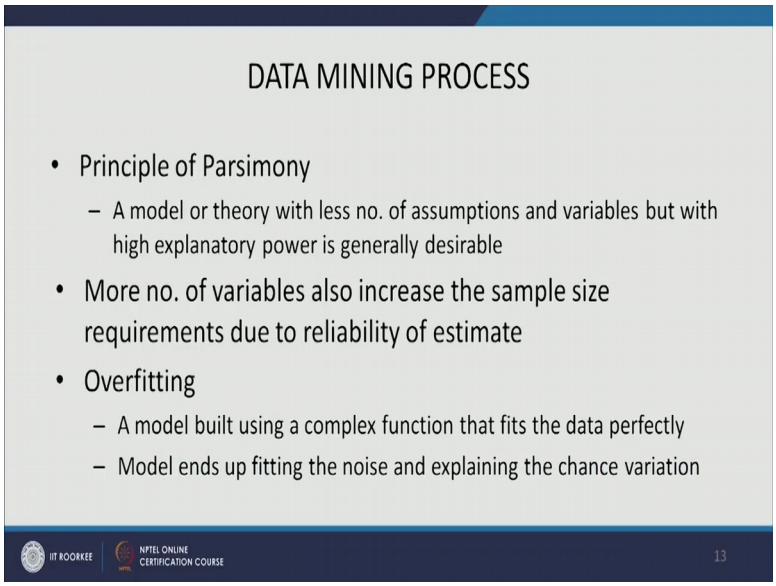
Now, there could be some statistical software which cannot use categorical variables expected in the label format. If you have format for example, if you have a variable like you know gender where you have male or female or M or F has been those are the labels that are they are in the data sets many a statistical software might not be able to use this particular variable the directly. So, therefore, dummy binary you know dummy coding might be required for these variables.

When we say dummy coding we actually create dummy binary variables. So, these are actually having 0s and 1s, 0 indicating the absence of a particular type and 1 indicating presence of a particular types. For example, if activity a status of individuals we have

data on activity status of individuals and has 4 (Refer Time: 23:29) and jointly exhaustive classes for as a student unemployed, employed and retired then in that case we can create different dummy variable wherever when a particular observation able to belongs to if its activity status belongs to a student, it builts have the value as 1 if it does not then it will have the value as 0, similarly for other observations and for other classes.

Now, if these types classes are usually exhaustive then we do not need to create all you know four dummy variables. So, there are 4 types we do not need to create 4 dummy variables because then being jointly exhaustive have been no 3 of them then the fourth one is already known. So, therefore, we need to create only three dummy variables.

(Refer Slide Time: 24:21)

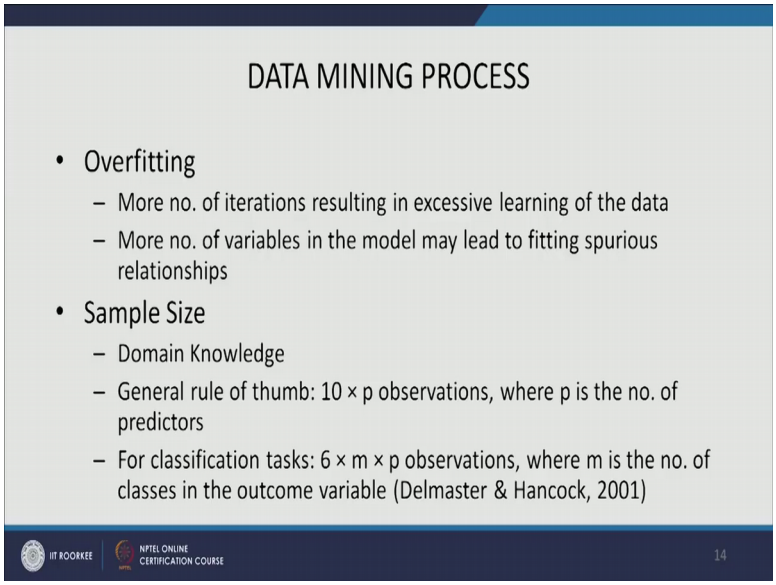


Now, another important concept is principle of parsimony. So, principle of parsimony is about when you know a model or theory with less number of assumptions and variables, but with high explanatory power is generally desirable. So, you are always looking to develop a model or theory where you make as few assumptions as possible and you include as few variables as possible and is still able to explain most of the phenomena you know higher explanatory power. Most of the cases you are able to explain and less number of your, you must be using less number of assumptions and variables. So, this is a desirable property and in this particular course we will be putting more impasses, impasses on this particular principle.

Now, another problem with you know more number of variable is that it is going to increase your sample size requirements because you are always looking for a reliable, you always want to estimate, you always want to compute a reliable estimate and if you have more number of variables then your sample size requirement will increase to have that higher reliability.

Now, another related problem is over fitting. So, what is over fitting? So, over fitting generally arises when you know a model is built using a complex function that fits the data perfectly. So, if your model is fitting the data perfectly then probably it might be over fitting the data. So, what actually happens is you in the over fitting you in the fitting the noise your model end up fitting the noise and explaining the chance variations. So, you would ideally you would actually look to avoid explaining chance variations because you are looking to understand the relationship which can then be used to predict the future values all right. So, therefore, over fitting is something that is not desirable. Over fitting can also arise due to more number of iterations, if we you do more number of iteration that can result in excessive learning of data. So, that can also lead to over fitting.

(Refer Slide Time: 26:32)



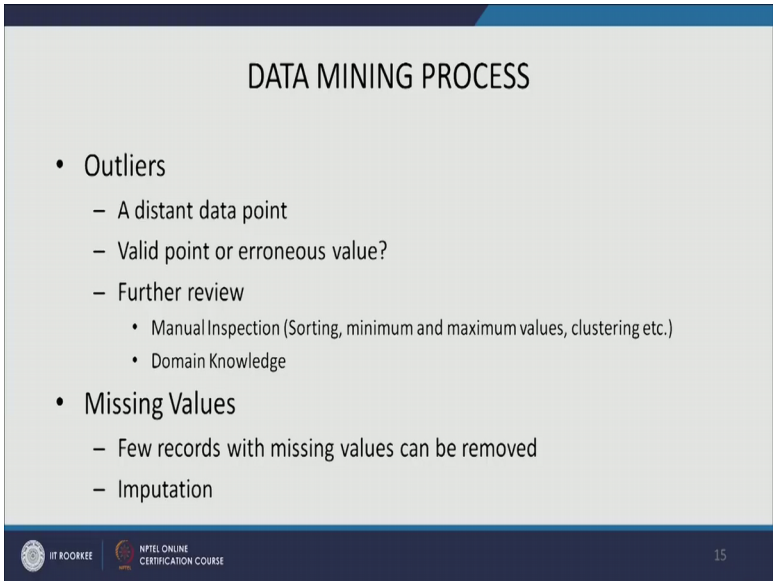
If you have more number of variables in model some of those you know variables might have a spurious relationship with your outcome variables all right. So, that can also lead to over fitting.

Now, the next concept is related to sample size. So, how many observation should actually be good enough for us to build, for you to build an accurate mode? So, domain knowledge is the you know is important here to understand a sample choice because as you do more and more modeling, more and more analytics you would be you know able to understand different phenomenas the construct concepts variables you will have a better hunch on or better rule of thumbs to understand how many observation will actually be we will to build a model for a particular analytics problems. So, domain knowledge is always going to be the crucial part.

We also have rule of thumbs for example, if you have p number of predictors 10 into p. So, for 10 observation for predictor can actually be a good enough rule of thumb to determine the sample size. Similarly for classification task also many researchers are suggested some rule of thumb for example, 6 into m into p observation, where m is the number of classes in the outcome variable and p is the number of predictors. So, that can actually help you you know determine a sample size.

Now, another relevant concept is, another element concept is outlier.

(Refer Slide Time: 28:19)



Now, what is an outlier? So, outlier can briefly be defined as a distance distant data point now important thing for us to understand is whether this distant data point is valid point or erroneous point because if a particular data point is distance from the majority of the values or very distance from the mean more than 3 d standard deviation away from the

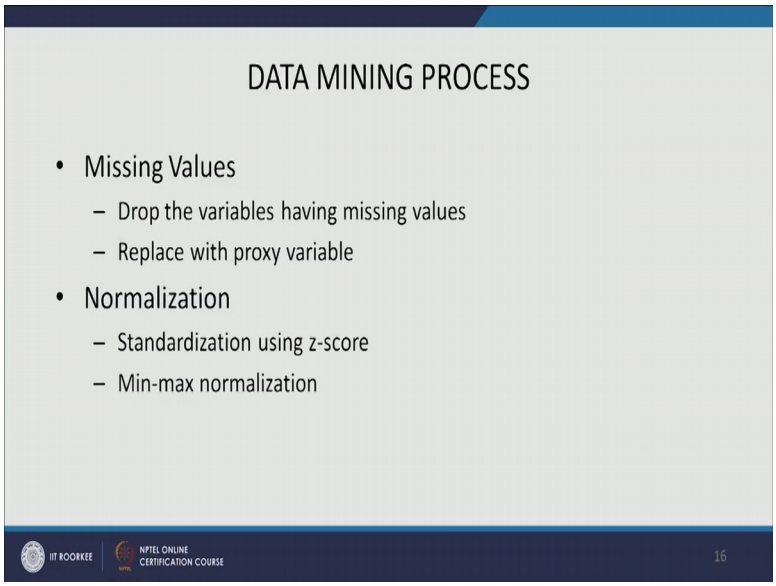
mean then it could be you know it could be due to human error or measurement error. So, we need to find out whether a particular data point is because of the human error or measurement error for example, at you know a temperature a room value of you know 100 or 150 for a room temperature or temperature in a city you know could be you know human error or measurement error. Sometimes there would be you know errors due to decimal points and related you know typing error and all that. So, we need to identify whether outlier whether it is a valid point or erroneous value.

So, how do we do that? So, generally you can do some manual inspection you can sort your values and find out if anything looks out of place you can also do I mean you can also look at the minimum and maximum value and form their you can try and identify whether they are outliers whether they are errors clustering can also help you. So, you can do clustering you could be able to see and whether a particular point is outline or not. Domain knowledge, domain knowledge also going to help about this.

Now, next related concept is missing values you might come across a data set which is very important for your particular you know your particular business analytics problems, but few records you know few records have missing values. So, if those records are few in numbers then probably you can remove those records and then go ahead with your analysis, but if the number of records are more than probably that can end up eliminating most of your observations. So, therefore, you need to handle those missing values.

Imputation is one way, so you impute those missing values with the average with the average value of that particular variable right so that could be one solution. You can also if that is also not, if that is also not desirable then another option is if you have many missing values then you can identify the variables where the missing values are there and if those variables are not very important for your analysis then probably you can think about you know dropping them. If those variables are important for your analysis then probably you can a look for proxy variable which is having less number of missing variables and replace that particular variable with their appropriate proxy variables.

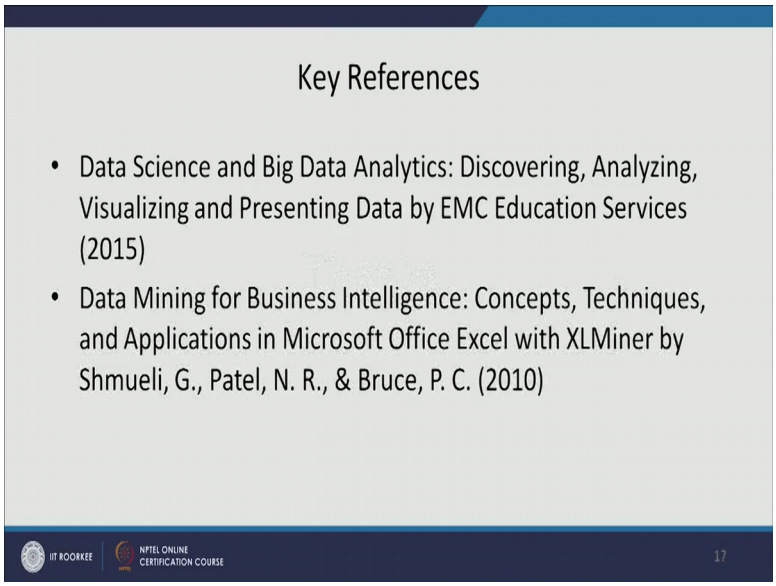
(Refer Slide Time: 31:15)



Another important concept is normalization. Many times in many techniques would actually require different variables you know you would require for you to normalize different variables. For example, distance based you know techniques you know clusterings and other where distance is an important computation important part of the process and, important part of the algorithm different you know variables having data and different units can actually create problem they can actually dominate the computation later with distance and we can influence the results.

So, in any scenarios normalization is the desired thing and the data mining process. So, there are two popular ways of normalization one is standardization using z-score where you subtract these values these value you know subtracted by mean and divided by standard deviation. Then there is another min max normalization where you subtract each value by the minimum value and then divide by the difference of max and min.

(Refer Slide Time: 32:35)



So, these are the key references.

Thank you.