Business Analytics & Data Mining Modelling Using R

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

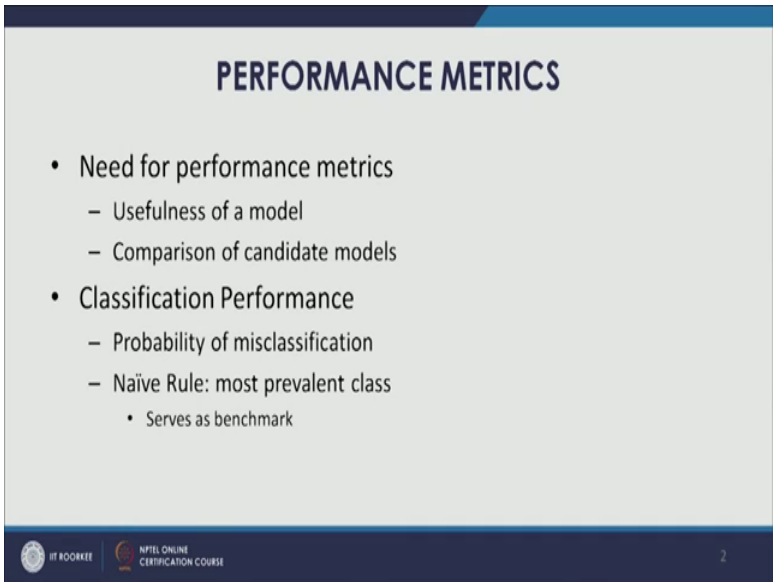
Lecture – 16
Performance Matrix-Part I

Welcome to the course business analytics and data mining modelling using R. So, in this lecture will start our discussion on our next module and next topic that is performance matrix for prediction and classifications? So, in this particular lecture we are going to focus on these two task prediction and classification and how different model whether they are prediction model or classification models, how their performance can be evaluated.

So, let us start our discussion. So, first of all we need to discuss why we need, why we require performance matrix. So, because different models, different methods can be applied to different prediction and classification task. So, therefore, how do we select the most useful or the best model from all those candidate models. So, for that we would require some matrix which can help us in deciding which can help us in making that decision finding the most useful model or finding the best performing model. So, comparison of candidate models can also be done, sometimes as we have talked about that it might not be several different methods just one method and different variants of it because of the different configuration or different options that we select for different variants.

For example, one simple example could be the same model regression model with different set of predictors could be run on the same data set. So, therefore, for the same method we will have a number of candidate models each having different set of predictors. So, therefore, it becomes important for us to have some matrix which can help us in deciding which one is the most useful model or the best performing model. So, in this particular lecture we are going to focus mainly on the classification and introduction and the relevant performance matrix, will start with classification.

(Refer Slide Time: 02:35)



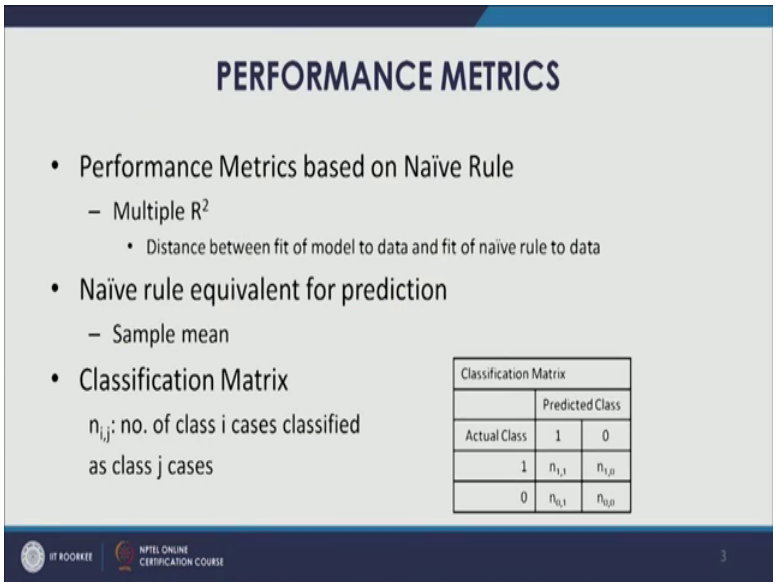
So, classification, generally this is the metric that we use probability of misclassification. So, the probability of misclassification is on the lower side. So, that can help us determine whether a particular model is useful or performing giving better performance or not. So, for different for different candidate models we can look at the probability of misclassification and then how that is how we can actually compare their performance or compare their usefulness for the task at hand, then naive rule for classification performance is could be the most prevalent class. So, for example, if we have 3, 4 classes for which we are trying to develop, for which we are trying to develop a classification trying to build a build a classification model.

So, any record, any record can be classified, any new record can be classified to the most prevalent record if it is classified to the most prevalent class then that is called naive rule. So, this serves as a benchmark. So, other models, other methods their performance can be compared with respect to this benchmark. So, we do not run a model, if we do not include the predictors information the relationship between predictors for our models, relationship between predictors and output variable in our model and just use the naive rule which is the any record irrespective of the you know information that is contained in predictors right we just assigned this particular record to the most prevalent class. So, for example, if there are 3 classes and the most prevalent one being class one this is having, let us see out of hundred observation 80 observation belong to this class and then the remaining 15 belong to class 2, out of remaining 20, 15 belong to class 2 and the 5

belong to the class 3. So, the most prevalent class being class 1, 80 records belong to this class. So, therefore, any new record can be easily classify to the class 1 and will also have a good amount of accuracy of based on this naive rule model 80 out of 100.

So, now any new method that we apply for this particular classification task classifying a record into these 3 classes, class 1, class 2, class 3 should actually perform better than the naive rule. So, naive rule serves as a benchmark and then we can look at the probability of misclassification and compare different candidate models to find out the most useful one or the best performing one, now there are few matrix which have been developed based on naive rules. So, therefore, these matrix actually use the naive rule and use it to build a metric that can be used to compare the usefulness of different model. So, one is multiple R square. So, what is multiple R square? More detail will discuss in coming lectures, but to give you a brief definition that multiple R square distance between fit off model to data and fit off naive rule to data. So, that is how that is why we are saying that this is metric based on naive rules.

(Refer Slide Time: 06:00)



So, distance between fit off model to data and fit off naive rule to data that is actually looked at in multiple R square.

If we want a naive rule kind of equivalent for prediction tasks, if we want a naive rule equivalent for a prediction problems, for prediction problems of predicting task sample mean could be the one. So, we are looking again if we are trying to predict a value for a

particular variable and we know, we are not interested in looking the information that is the that can come from predictors right we are ignoring the predictors information then simply the mean of a mean value of that particular output variable can be taken, can be taken and for any new record that mean value can be assigned to that record. So, that could be the naive rule equivalent for prediction tasks because the mean value this is this is the centralizing value for a particular variable and therefore, any new variable we can expect it to lie around you know that mean value. So, the error would be minimum if we do not use any other predator information the simple, sample mean could be a new rule equivalent for prediction tasks.

Now, let us come back to our discussion on classification. So, classification on matrix is generally used to compute a different performance matrix, different performance metric for classification tasks. So, you can see how a classification matrix is displayed, you can see here. So, generally will have predicted class, so if we are if we. So, in this particular case we are talking about 2 class scenarios where one class is represented by 1 and the other class is represented by 0 value, similarly actual class represented by 1 and 0.

So, that we have predicted class and actual class, class and we have different numbers. So, this is based on once the models have one once different models have been applied. So, for a particular model for a particular model n i j represents the number of class i cases, classified as class j cases. So, we look at and n 1 1. So, this is the number of a class 1 cases, classified as class 1 cases, this is the true classification right. If we look at the next number this is number of class 1 cases we classified as class 0 cases. So, this is this is the incorrect classification, similarly we look at the second class that is class 0. So, the this particular value and n 0 1 is number of class 0 cases classified as a class 1 cases. So, this is also incorrect classification on next value is n 0 0 this is number of class 0 cases classified as class 0 cases. So, this is the correct classification.

So, if we look at the diagonal values n 1 1 and n 0 0. So, these 2 values are actually the correct classification by the models. So, these numbers so these some of these numbers represent the correct classification and will look at the off diagonal values that is n 0 1 and n 1 0 they represent the incorrect classification. So, some matrix can be computed based on this classification matrix. So, generally most of the techniques most of the methods that we apply that we use for a classification task they generally you know, they

generally result in classification matrix. So, finally, with this classification matrix is then used to compute different matrix and then evaluate the performance.

Let us another important points about classification performance is that while we build our classification model on training partition it is the so therefore, the while we are building our model on training partition. So, the model will fit the data little bit more and therefore, as we have been talking about it is not recommended to test the performance of the model on the same partition. So, therefore, it is the validation partition classification matrix, that is generally used to just the performance of a classifier performance of a model for classification tasks.

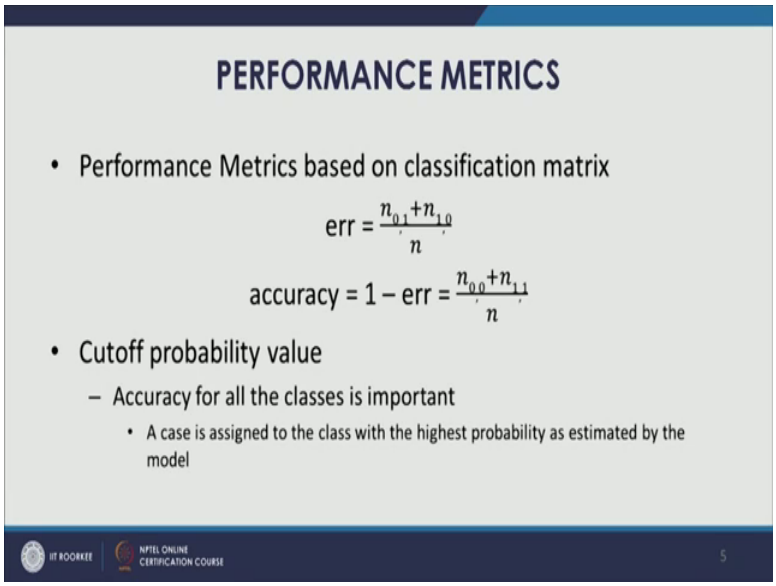
Now, we can still develop a training classification matrix, but it could be used to compare with the validation partition classification matrix to detect over fitting. For example it is expected that model would perform slightly poorly for the validation partition because the model was built on training partition. So, it will fit the training a partition data more accurately. So, therefore, the performance might go down for the newer partition that is valuation partition; however, if there is too much of gap between the performance numbers of validation partition classification matrix and training class partition classification matrix that might indicate over fitting, that the model has over fitted the training partition and that is why the performance on validation partition has come down much further. So, that is how that is one way to detect over fitting if the numbers for if the numbers for validation partition cosmetic classification matrix is on the lower side, but not that much there is a small gap then probably the model is stable and performing the develop model is performing well might perform well on the new built data as well.

Now, there are some performance matrix as we were talking about which are based on classification matrix. So, these are as simple as misclassification rate or error and accuracy. So, if we go back to the classification matrix that we discussed just a bit before, you can look at the off diagonal values n 0 1 and n 1 0. So, these are the misclassified number of cases. So, misclassification rate or error would be the proportion of you know proportion of total number of observation which are misclassified. So, these are going to be n 0 1 and n 1 0, if we look for the another matrix that is called accuracy. So, therefore, the classes as the observation classified into their actual classes in that case we go back to the matrix we would see n 1 1 and n 0 0 are these 2 diagonal values they

represent the accuracy, proportion of these numbers out of total observation represent the accuracy and it reminds the usefulness of a particular model.

So, misclassification rate or error or accuracy numbers can be used to compare the performance of different candidate models.

(Refer Slide Time: 13:32)



So, these are the formulas for these 2 matrix, 1 matrix based on classification matrix you can see error is computed as n 0 1. So, plus n 1 0 divided by total number of observation that is n and accuracy is one minus error that is n 0 0 plus n 1 1 divided by total number of observation n. So, these 2 matrix can actually be used to compare the performance of candidate models, another important concept related to performance matrix is the cut off probability value.

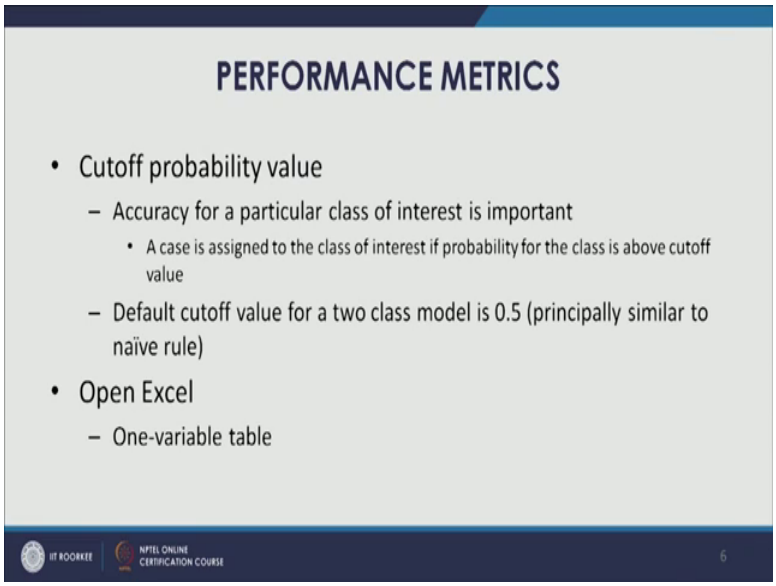
So, cut off probability value is. So, we talked about the earlier we talked about the probability value, earlier we talked about the probability of misclassification. So, let us come back to the same discussion. So, how do we, how do we assign different cases, how do we assign different cases to different classes. So, generally the data mining methods or statistical method, generally they compute using the model they compute the probabilities values for different record. So, if there are 2 classes for example, class 1 and class 0 for each record will have probability value of that record belonging to class 1 and probability value of that record belong to class 0.

So, how do we, how do we assign that record to a particular class. So, cut off probability value is the 1 that is used to do that to perform that that assignment. So, there are different scenarios how this cut off probability value can be used. So, first one is when accuracy for all the classes is important and that means, we want to correctly classify all the cases irrespective of the class; that means, they could belong to class one and class 0, but we are trying to classify them if the record belongs to if the actual class of a particular record is class 1.

We want to classify it to class 1 and if the actual class of a record is class 0 through our model we want to classify as class 0, if that is the scenario if that is the that is what we want then the particular case or observation can be assigned to the class with the highest probability as estimated by the model. So, as we talked about for every case or observation will have number of probability values for each class. So, for 2 class case class 1 will have a probability value for class 0 also will have a probability value. Will compare them and the, and the class where the probability value is on is more than the other value same value for the other class then the case would be assigned to the class having the higher probability value.

Now, in some other scenarios accuracy for a particular class of interests is important. So, there could be a particular class where we might be interested in identifying cases belonging to that class a bit more. So, we would be interesting in identifying ones little bit more even if it comes at a higher misclassification records for 0, class 0 records belonging to class 0. So, in those situations the assignment would happen in a different fashion. So, we would not be assigning depending on the higher probability value. So, what we generally do is case is assigned to the class of interest, if probability for the classes have a cut off value.

(Refer Slide Time: 17:20)

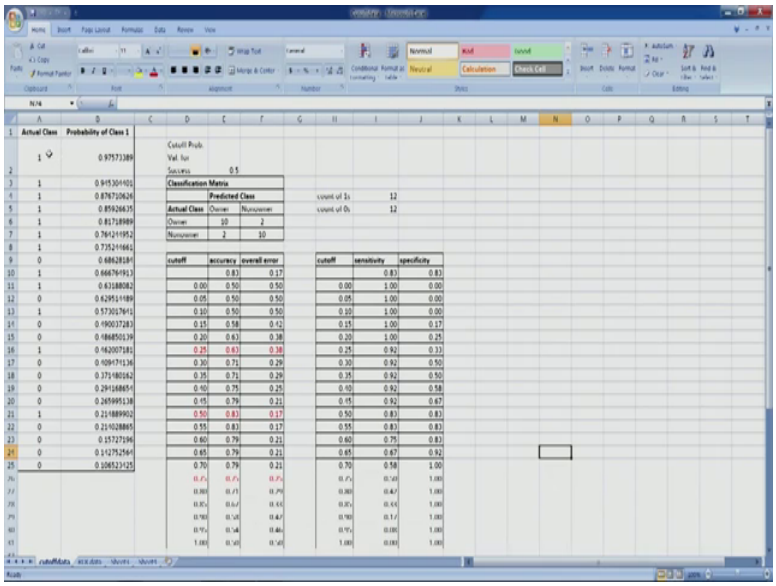


So, will define a cut off value in such cases and if the value for class of interest is higher than the cut off value that particular case or observation would be assigned to that class of interest otherwise it would be assigned to the other class.

So, if we talk about two class model, the default cut off value is generally 0.5 that is principally similar to the naive rule because if there are 2 classes if a particular probability value is more than 0.5. So, therefore, that is the more prevalent class. So, that is how the assignment is going to happen, if we look at the other scenario when accuracy for all the classes is important even there when a record is classified to the highest probability value class there also we can see the application of naive rule was higher, higher probability meaning highest probability meaning that is the most that is in a way more prevalent class. So, therefore, that that particular record is being assigned to that particular class so the idea is similar to naive rule when we talk about the assignment based on the probability values.

So, we will do an exercise using excel. So, let us open this particular file. So, what data we have is.

(Refer Slide Time: 19:04)



So, the here are 24 observation as you can see row number 2 to row number 25, we have 24 observation and for each observation we have the actual class whether the particular observation or case belong to class 1 or class 0. So, you can see that is appropriately indicated here 1 or 0 for each observation and we also have probability of that particular class belonging to class 1. So, that probability value is given there, probability of a class belonging to class 0 would be one minus this probability because we are discussing 2 class scenario. So, these probability values are here also and if you see they are arranged in the higher probability to lower low probability sequence they are arranged higher probability to low probability sequence.

You can see if we cut off probability value for success the default value that is 0.5 as we discussed before. So, we look at the actual class 1 it is going to be correctly classified as class one because the probability value for this particular case is more than 0.5 similarly for second observation this is also going to be correctly classified having more than 0.5 value. So, same is for this case if we keep going on in this fashion will these 2 record or observation number 8. So, there you will see that even though the cut off probability value is 0.686 which is more than 0.5 the actual classes 0. So, as per the probability value it would be classified as 1, but the actual class is 0. So, therefore, this will we and this is going to be an incorrect classification.

If we move further again we get the correct classification, how more than better more than 0.5 value and correctly classified, if we go further again we encounter another misclassification where the value is for the observation is more than 0.5, but incorrectly classified. Similarly now further as we go down all probability values start becoming less than 0.5 and the classification also start becoming 0. So, in this case observation number 13, case number 13, valuing 0.49 and the observation being correctly classified as class 0. Similarly, if we go down there are few misclassification for example, this one record number 15, row number 16 this is .46 value and this is incorrectly classified as this is going to be incorrectly classified as 0, but the actual class is 1.

So, we can say if we want to count the number of ones and 0s in this particular example that we have, number of ones are 12 and number of 0s are 12 and if we look at the based on this class cut off probability 0.5 if we construct the classification matrix. So, these are the values. So, classification matrix we can see that 10 observations are which are, which belong to the for example, owner for example, their owner is 1, non owner is 0. So, 10 observation are correctly classified 2 observation which are actual owners, they are incorrectly classified, if we look at non owner class. So, 2 observation which were actually non owner they were incorrectly classified as owner and 10 observation which were actually non owner correctly classified as non owner.

So, these are the numbers and the accuracy values can be computed you can see, you can see here the diagonal values diagonal values e 6 and f 7 and they are divided by the total number of observations and we get the accuracy similarly for error also. So, we have created so here in this particular case we have created one variable table in excel and for these value range from 0 to 1 at every 0.05 interval. So, as the cut off value changes from 0 to 1, we have the accuracy and overall number. If we focus on these values 0.25 we will have the accuracy number as 0.63 and 0.38, if we go back to the default value 0.5, 0.83. So, you would see that at the default cut off value the accuracy is much better for 0.25 it was 0.63 for 0.5 it is actually 0.83. So, default cut off value accuracy numbers are much better.

If the cut off value was 0.75 then you would see again the accuracy is comes down from the case of 0.5, now is 0.75 from 0.38. So, whether we move, whether we reduce the cut off value from 0.5 to 0.25 the accuracy goes down or whether we increase the cut off value from 0.5 to 0.75 again the accuracy goes down. Now, we look at the if we look at

the classification matrix so you would see right, now the predicted we look at the predicted class right. Now 12 are being predicted as owner and 12 are being predicted as non owner when the cut off value is 0.5, if we change this cut off value to point from 0.5 to 0.25 you would see the classification matrix has changed because of the appropriate formula is being used in those cells.

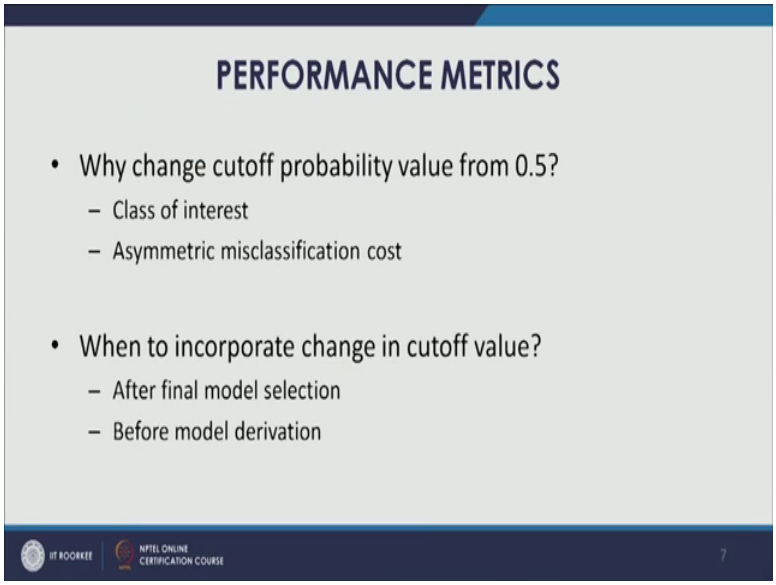
So, you would see now in the protected class 19 observation have been classified as owner and only 5 observation has been classified as non owner, this is mainly because of the lower cut off value 0.25. So, therefore, more number of observations are being you know classified as ones and therefore, we see more number of owners 19. So, this jump from 12 to 19 and that is why there is a drop in the accuracy value, similarly if we change the cut off value to 0.75 you would see the numbers have changed in the classification appropriately and now you would see that there are more non owners being more observation being classified as non owners we would see that only 6 observations have been classified as owner and 18 observations have been classified as non owner.

So, you would see that because the cut off value is more than 0.5 on the higher side. So, therefore, fewer observations are being classified as ones and more observations are being classified as 0. So, let us go back to our discussion. So, if we move away from the cut off value of 0.5 we move to a 0.25 or 0.75 then in each case the accuracy was down. So, in each case the accuracy goes down. So, why change this cut off probability value from 0.5, there could be two regions one is we have a special class of interest and therefore, we are interested in identifying those rare cases all those cases belonging to a special class of interest a little bit more, even if it comes at the misclassification of higher misclassification weight for other class, similarly another region could be asymmetrical misclassification cost.

So, if one particular class the cost of misclassification for that particular cost would be much higher than the cost of misclassification for an for some other class. So, therefore, because of this asymmetric misclassification cost also we would be interested in finding out more of the that more of the observation belonging to class having higher misclassification cost and that would require us to move away from our default cut off probability value of 0.5.

So, next question would be when to incorporate change in cut off value, when do we change.

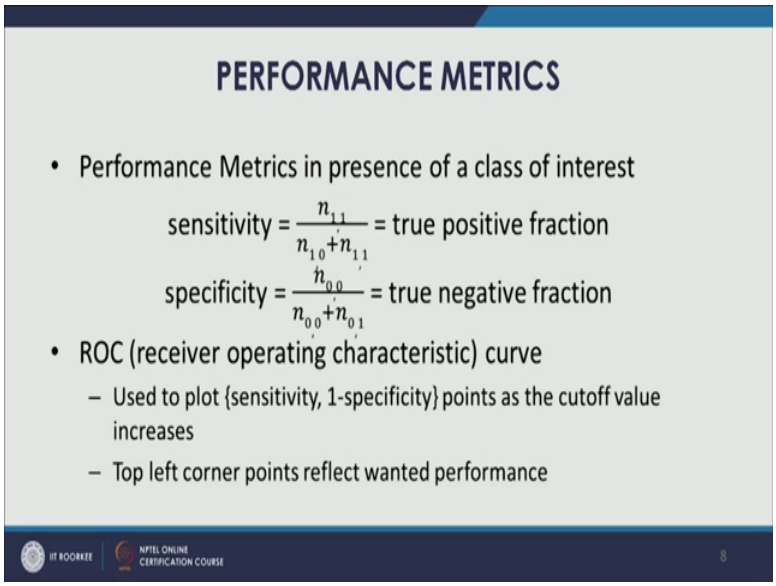
(Refer Slide Time: 27:56)



So, the so for example, if the example that we the exercise that we just went through in excel was be already had we had already run the model, we had the actual 1s and 0, but we also had the probability values at as estimated by the model. So, you would see that the exercise of changing the cut off value from 0.5 to 0.25 or 0.75 or other numbers through one variable table in excel that was done after the model has been selected and we had the probabilities value and then we were trying to see how the results could change if we change the cut off value.

So, one situation would be after final model selection, we can incorporate change in cut off value because will have the probability values most of the techniques they pro they estimate probabilities values also. So, therefore, it is easier for us to change the cut off value and see how the results are changing; now another situation could be before model derivation. So, we can incorporate the misclassification cost that could be there at the model derivation during the model derivation steps itself and that would ultimately determine the results. Now, when we have a special class of interest the performance metric like accuracy and error might not be useful. So, few other metrics are popular when we are specifically interested in a particular rare class or a particular class of interest. So, these metrics are sensitivity and specificity.

(Refer Slide Time: 29:47)



So, will discuss them in more detail in the next lecture so will stop here.

Thank you.