Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

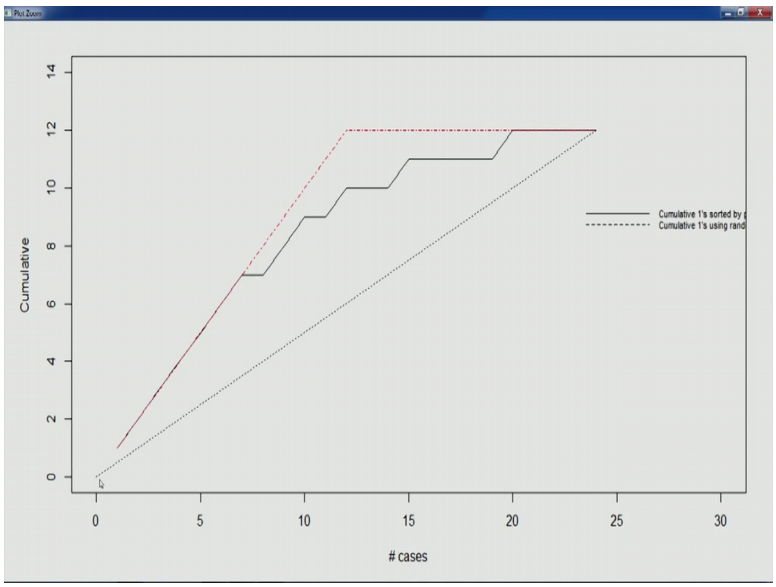
Lecture – 18
Performance Matrix – Part III ROC Curve

Welcome to the course Business Analytics and Data Mining Modeling Using R. So, in the previous session we were discussing Performance Matrix and in particular we were we had created a cumulative lift curve or gains chart.

So, this was the chart that we had created in the previous lecture. So, as you can see on x axis we have the number of cases. So, cases at cases are increasing from 0 records to 5 records on x axis and 10 records and then 15 20 in this fashion. So, as the number of cases increasing what is the performance of the selected model in the y axis we have cumulative actual class.

So, cumulative actual class being identified by the model as we increase the number of cases, now the dotted line is the reference line where we have the cumulative ones as would be identified using a random selection approach, which would mostly rely on the probability value of a particular record belonging to a particular class right. So, when we start when we have 0 observation this particular reference line will you start from 0 and 0 no records.

(Refer Slide Time: 01:40)



And no identification or classification as we move to all the records that is 24 observations in this case you would see in the data set we had 12 we had 12 owners and 12 non owners.

So, therefore, the and the class are being interest being owner. So, number of records belonging to the owner class are 12. So, the probability of a particular record belonging to owner classes 12 divided by 24 that is 0.5. So, you would see data the last point coordinates being 24 and 12 if you look at so this being the reference line. So, this being the average case now how our model is performing you can see the cumulative actual class values and how this particular line in black color stepwise line in black color is the performance of our model, as the number of cases increase we can see the gap or separation between the this reference line and the actual line is increasing; that means, model it starts to perform better in comparison to this random selection case.

So, as we move further there are slight blip blips in between where we take these steps and then again the performance improves, where we see the slide slide blips our horizontal lines these are actually misclassification. So, our model has misclassified a particular observation. So, class of interest being class one and that is repre being represented in the y axis therefore, a class a an observation has been identified probably as class 1 which was actually class 0 right.

So, if we had a model which a would explain all the observations perfectly. So, each record to it is own class then that would that model would actually be represented by red dotted line.

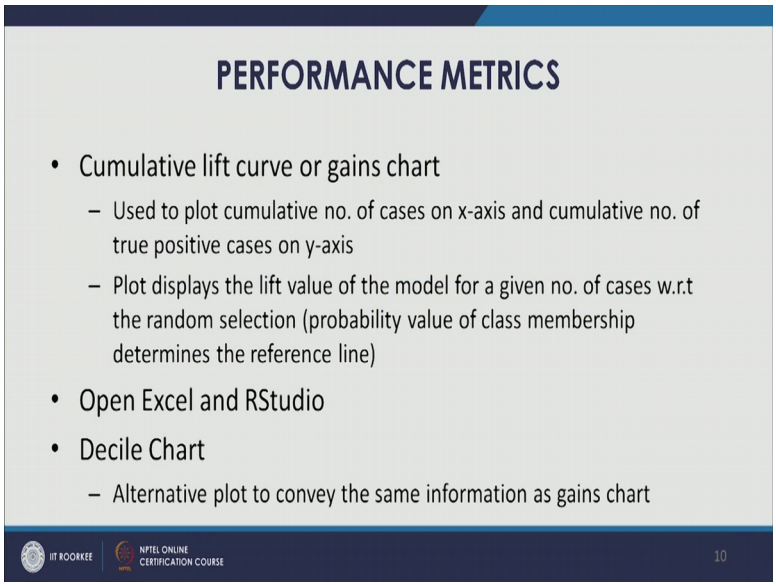
So, you would see this particular model as we move along right. Now this particular data that we have plotted in cumulative lift curve or gains chart is actually reflecting the rank ordering as well. So, you would see that all the this by this red model red dotted a model that the perfect model will represent will correctly classify all the classes belonging to owner class you can see this particular point is 12 cases.

So, all the cases have been correctly identified and then you would see the horizontal line because now all the cases belonging to non-owner classes are being identified. So, this par particular model is will go in this particular fashion our model they are going to be few misclassifications. So, therefore, our model will deviate from this particular red line,

but if we compare with the reference line there is much more better performance by the model.

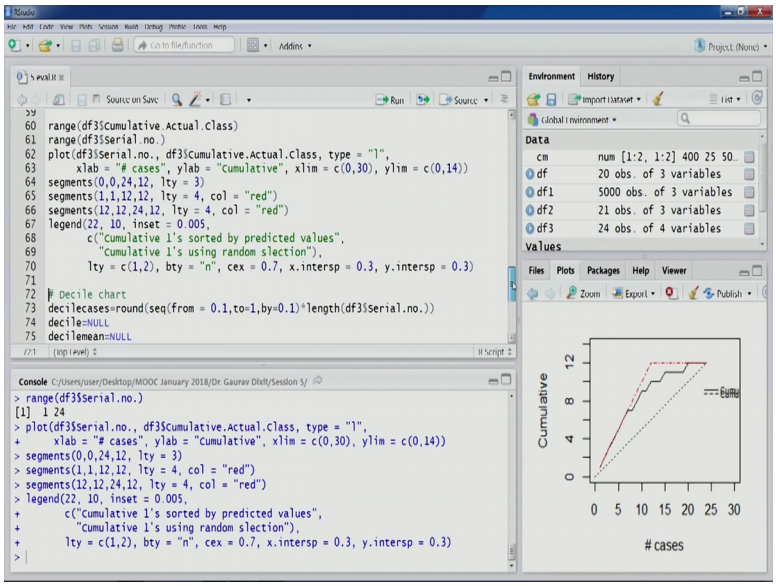
Now, let us go back to our discussion now what we have discussed in cumulative lift curve or gains chart.

(Refer Slide Time: 05:01)



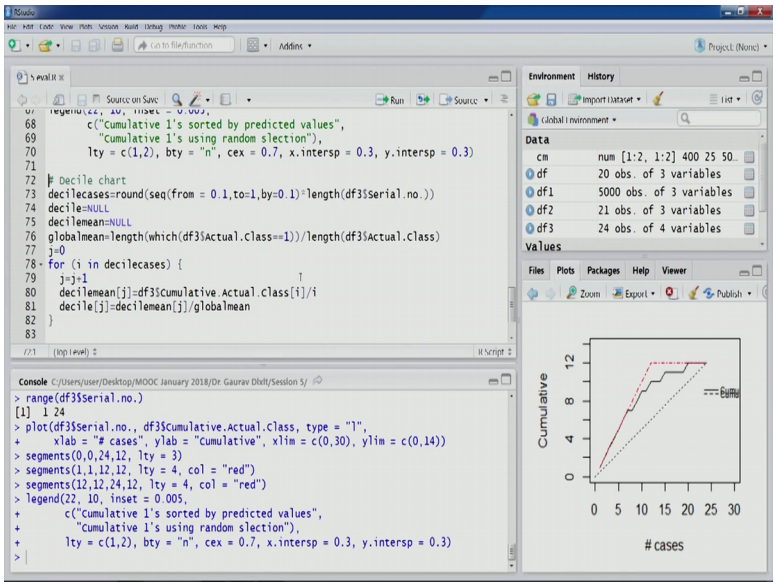
In a way helps us in determining how much improvement our model is providing if we had not used any model at all what was the case, which was represented by reference line and the gains that the lift from that no model scenario that lift that has been given by the model that we could see in this chart. Now the same information that we just plotted that we had seen and plotted through gains chart can also be done through decile chart now this is alternative plot to convey the same information as gains chart right.

(Refer Slide Time: 05:45)



So, let us open our studio and will generate decile chart so before generating decile chart.

(Refer Slide Time: 05:49)



We would need to create a few variables horizontal, decile chart as we will see that we create different deciles which are nothing, but buyers which represent the cumulative.

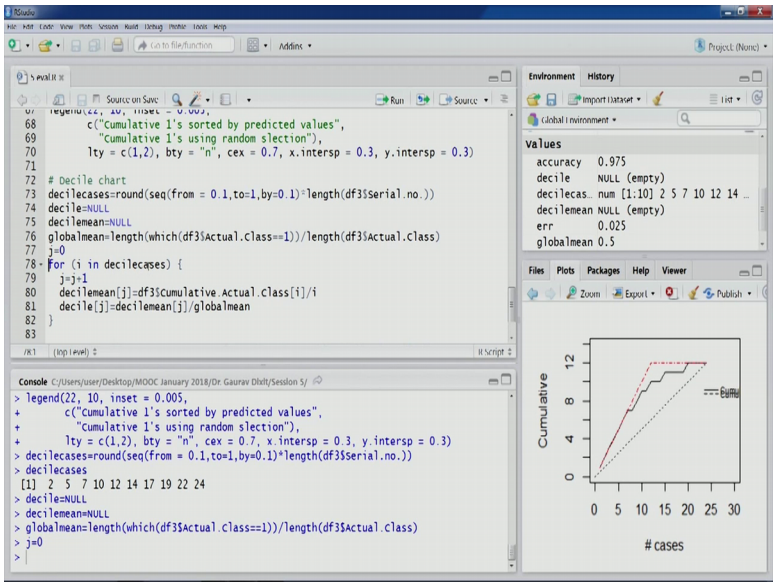
But number of cases in percentage terms 10 percent of cases and what was the bar size then 20 percent top of the cases and what was the buyers size and this all and then 30 percent of the cases and the respective bar plot.

So, in this fashion we plot for all the cases. So, that is all 100 percent cases. So, if we have gaps for 10 percent 20 percent 30 percents so will have 10 such bar. So, for each bar we would try to compare each bar would reflect the same kind of information which was reflected by gain chart the lift that is given by the model.

So, I will see how we can generate this. So, as we saw in the gains chart the comparison was with respect to reference line. So, in this case on y axis will have decile mean divided by global mean. So, for every decile or bar will create the mean for all those values and that would be divided by the global mean. So, global mean here being this value, which is actually the number of cases belonging to actual class 1 divided by number of total number of total cases.

So, this particular value will give us the number of will give us the global mean the average case. So, let us first create this variable decile cases. So, this is nothing, but variable which is representing the number of observation that are going to be there for each decile or each bar.

(Refer Slide Time: 08:02).



So, you first decile will show some stat or some show bar about 2 observation, than 5 observation, 7 observation, 10 observation in this fashion.

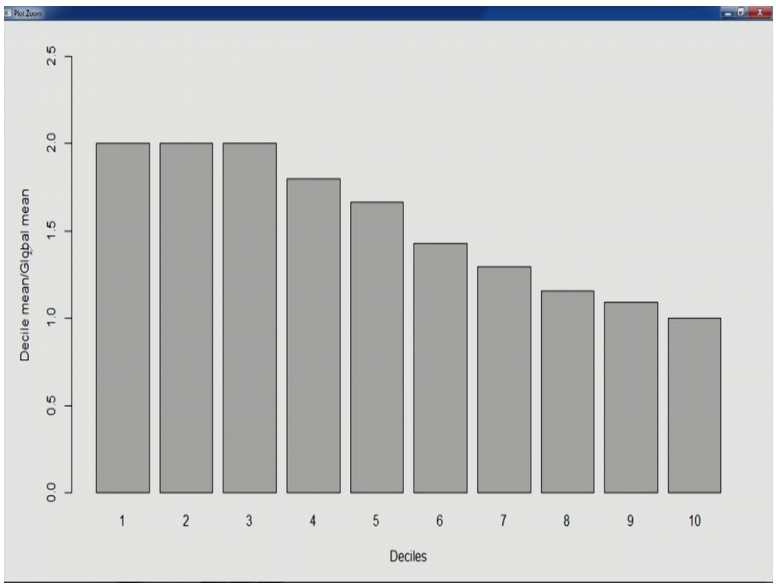
So, do observation actually representing 10 percent of the total observation in this case 24. So, it has been rounded to 2 then 5 this has been this is 24 percent of the total

observation that is 48 rounded to 5 observation and in this fashion for different number of observations different deciles will have different number of observations. So, let us create few more variables. So, that we are able to compute this decile mean for each decile global mean has also been computed as you can see that this value is 0.5. Now let us also compute this particular counter and is like this particular counter.

Now in this particular for loop what we have is we are running a counter I for all the decile cases that is in this particular cases there are 10 such deciles. So, the loop is going to be run on 10 times and you would see in the next second line; in the for loop decile mean for every decile is being computed and you can see that cumulative actual class is you know is being divided by the counter value and that that giving us the decile mean for that particular, because the high will have actually the number of decile cases.

So, far a particular bar the number of cases so, the cumulative actual class is being divided by the number of cases for that particular decile. So, will have decile mean and then decile value for is being computed by dividing the decile mean why global mean. Let us execute this for loop will have decile numbers. Now will create a bar plot will actually create this decile. So, let us compute this you would see a decile chart has been generated.

(Refer Slide Time: 10:17)



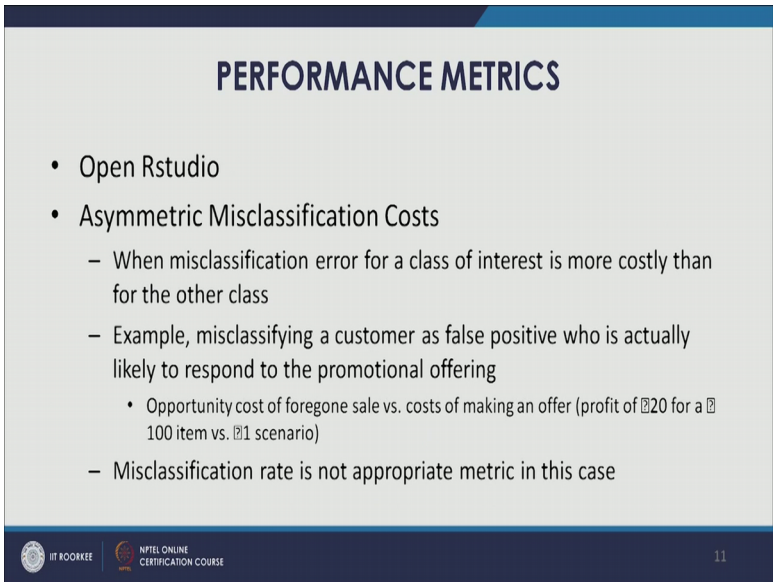
In on the y axis we have decile mean divided by global mean and you would see that this particular and the on the on the x axis we have deciles. So, for decile rep representing 10

percent of the cases and these 10 percent of cases are actually from that ordering rank ordering that we have done based on the estimated probability numbers.

So, these are the most probable ones coming first then 10 percent most probable ones coming first, just like the cumulative lift chart, they are also the most probable ones were coming first. Similarly second decile most probable ones 20 percent of such cases coming first and then 30 percent, because as we talked about in a business scenario in a business context we would be interested in identifying for example, in a case of promotional offer we would be interested in identifying the customers who are most likely to respond to a particular promotional offer therefore, most likely once we would like to mail first.

So, these 2 particular charts that we talked about are actually indicating the performance of our model in that sense in rank ordering sense. So, you would see that lift for decile 1 is actually 2 in comparison to the average case for 20 is for decile 2 that is for 20 percent cases again the similar number for 3 also the decile number 3 also similar number, as we move towards right side of x axis we would see that decile value keeps on decreasing, because as we go along and try to classify more number of cases the lift values goes down. The same thing is same thing was reflected in the this particular gains chart that we had, we compare from here the random case this this average case and we see the separation we would see as we move further from this point onwards, you would see this both these lines are closing right the same thing is reflected in the decile chart as well.

(Refer Slide Time: 12:43)



So, let us go back to our discussion on performance metrics. So, next particular concept that we are going to talk about is asymmetric misclassification cost. So, we talked about that when we would require to change the cut off value we talked about different scenarios, when we have a particular class of interest then other scenario going that when asymmetric classification cost are involved. So, what is this particular concept? So, when misclassification error for a class of interest is more costly than for other class who sometimes it might.

So, happen then that if a particular class of interest if there is some misclassification error it is more costly this generally happens in a business context for example, and misclassifying a customer as false positive who is actually likely to respond to the promotional offering. So, for example, if we were able to make a promotional offer and that if that offer was accepted by the customer and he purchases that particular item then we might end up making some profit. So, if we are not able to correctly identify or classify that particular customer as the responder of the offer then we is stand to lose that particular profit that we could have made. So, this is opportunity cost of foregone sale.

So, if we are not able to correctly classify a responder then we will lose on this particular opportunity cost opportunity cost of a forgone sale if we are. So, this particular cost has to be compared with the cost of making an offer if the particular customer is not a

responder then the cost of making an offer would be much lesser than the opportunity cost of forgone sale.

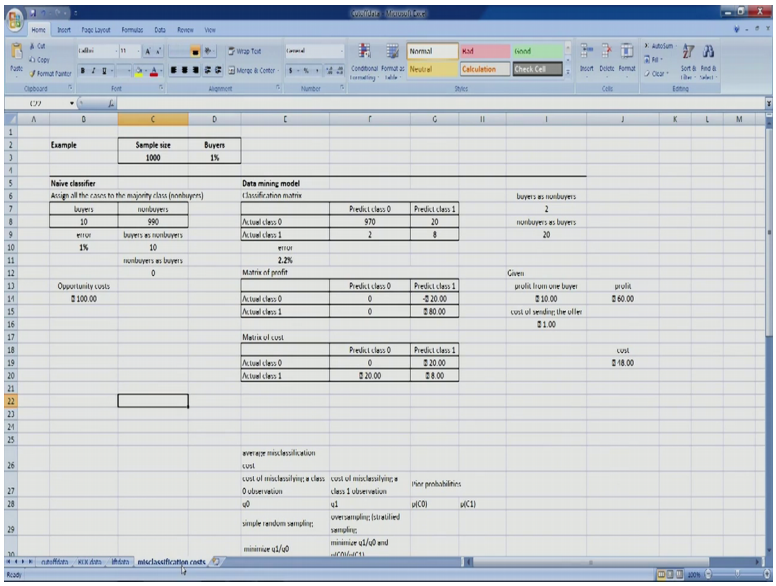
So, it is important for us to identify a particular class the class of interest because if we are not able to do. So, we would that is going to cost us more. So, in this particular scenario the misclassification rate metric that is generally applicable is not appropriate reason being there are different cost in different cost associated with misclassification of different classes, other consideration that could be there for example, it might not just be the opportunity cost of forgone sale versus the cost of making an offer there are some cost of analyzing data also.

For example if we built a data mining model that would require us to have the data set in the first place it might be in the form of data warehouses data marts and then the finally, data would be extracted for the analytical problem and therefore, will develop a classification model a classifier and that will help us identify. So, all that is going to incur a cost so will have to we should be incorporating that cost as well if through this model we identify a particular class one member then we should be incorporating this cost as well. So, we need to look at the actual net value impact per record.

So, if we are able to correctly identify with the help of our model data mining model then we have to incorporate the cost of analyzing data as well. So, actual net value impact per record would be much better. So, eventually our goal when asymmetric misclassification cost are there our goal would not be minimizing the overall error or maximizing the overall accuracy rather it would be minimization of cost or maximization of profits.

Now, while we do this how to improve actual classification by incorporating asymmetric misclassification and cost how that can be done. So, will see that, but before let us go through an example.

(Refer Slide Time: 17:00)



So, let us open excel file to understand misclassification cost further before we discuss how we can incorporate miss misclassification cost to improve our modeling.

So, let us say we have this example we have 1000 observations this is our sample size and one percent of this particular sample they belong to the buyers category. So, that would be 10 buyers and then remaining would be non-buyers. So, buyers would be 10 and the remaining would be non-buyers. So, if we use the name classifier we do not build a data mining model then what we will do following name classifier is assign all the cases to the majority class.

So, for example, in this case we look at the sample then there are 10 buyers. So, all those buyers would also be classified as non-buyers that would lead to 1 percent error. So, will have 99 percent accuracy or 1 percent error that seems to be a good model, but this is of no practical use reason being that we are not able to make any money out of it because all the customers have been classified as non-buyers.

So, the opportunity cost for this particular model the name classifier is going to be 100 rupees given that profit from one buyers 10 rupees and cost of sending the offer is 1 rupees.

So, if we look at I have already done the formulas here. So, you can see here that the numbers are buyers multiplied by the profit from one while that could be there that gives

us the opportunity cost numbers that is 100 rupees. So, this is the cost if we follow the naïve classifier. Let us say we had a data mining model and using this particular model we generated this classification matrix. So, this particular model correctly classified 970 class 0 as class 0 members.

But 20 were misclassified so 20 of class 0 members were misclassified as class 1 to class 1 members were misclassified as class 0 and 8 class one members were correctly classified as class 1 members. So, that gives us in terms of misclassification 2 buyers has been misclassified as non-buyers and 20 non buyers have been misclassified as buyers.

So, that gives us an error of 2.2 percent you can look at this particular value and see here that diagonal values of diagonal values and divided by all the observation that gives us the error that is 2.2 percent, if we compare this particular scenario with the naive classifier the error is on the higher side.

But if we look at the classification matrix now we are able to identify correctly identify 8 buyers. So, that will give us some profit so how we can understand that. So, one way to create one way is to create a matrix or profit. So, here will look at depending on the results of classification model, how we can make profit from that. So, because we were able to identify 20 and 828 members as belonging to class 1 therefore, from these, but only 8 of them were correctly classified as 1. So, from these 8 members from these 8 members will have 80 rupees 10 from each while and 20 which were you know which were classified as class 1, but they are actually class 0 members. So, will have 1 rupee cost of sending the offer, that is the cost, that is minus reflected through minus sign minus 20 so overall will have a profit of 60 rupees.

So, this is a from the profit perspective we can look at the same situation from the cost perspective. So, cost of sending the offer is 1 rupee and then the profit from 1 buyer is 10. So, will look at the number of number of customers who have been incorrectly number of class 1 customers who have been incorrectly classified as class 0 so, these are there are 2 such customers and profit that we could have made from them is 10 so, this is 20.

Now, other members which have been classified as 1 20 misclassified and 8 correctly classified. So, 20 and 8 is the value. So, we look at from the cost perspective total cost is going to be 48. Now we can either maximize this particular figure we can either

maximize. So, our naive word would be maximization of this profit or minimization of this cost. So, let us go back to our slides.

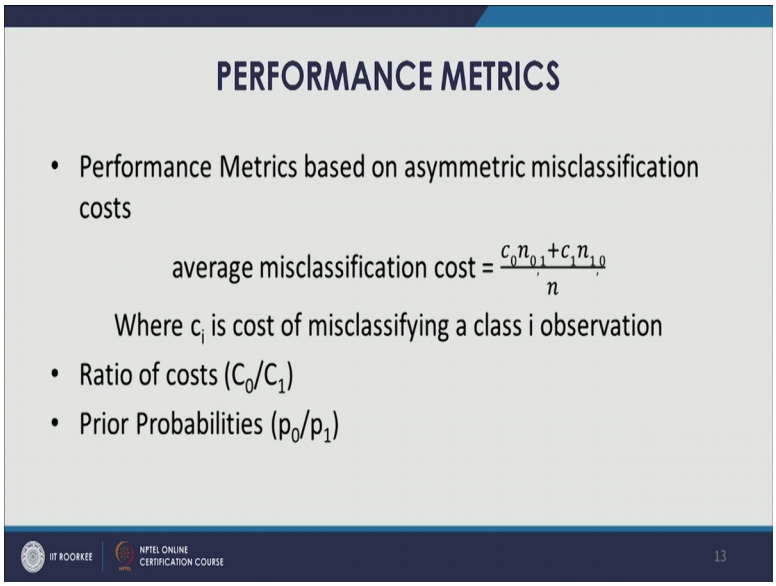
Now, as we talked about that the actual classification would not be improved by what we discussed just now whether by minimization of cost or maximization of profits through these new goals. So, how do we improve actual classification one way is change the rules of classification for example, in the previous lectures we were talking about changing the cut off value from 0.5 we can either increase or decrease? So, that that could be done

So, that is that is one example.

So, what we can do here is we can create a new performance metric. So, this is actually the average misclassification cost. So, here we can look at this particular value C I which is the cost of misclassifying a class 1 observation. So, in this case we have 2 classes. So, C 0 is the cost of misclassifying a class 0 observation and C minus is the cost of misclassifying a class 1 observation. So, average misclassification cost this naive metric can be computed using C 0 into N 0 1 that is number observation number of observation belonging to class 0 classified as misclassified as one then plus cost of misclassifying a class one observation that is c 1 into number of observation or belonging to class 1 misclassified a 0. So, this is misclassifications all the misclassification and the misclassification cost have been computed in the numerator divided by the total number of observation.

So, we can actually look to minimize this average misclassification cost. So, that is how we can actually incorporate. So, this could be our change this could be new matrix where we can look to minimize this average misclassification cost and thereby changing our classification rules and improve the classifications.

(Refer Slide Time: 24:48)



Now, for this we would have to estimate C 0 and C 1 values right so, but if we look at this particular formula if we divide we take c 1 outside of this will have a ratio C 0 divided by C 1. So, other term would actually be the constant value. So, therefore, if we do not need to find out the actual cost of misclassification C 0 C 1 if there are more than 2 classes this could be a costly process.

So, we do not need to do that if we are able to if we are able to find out if we are able to understand the ratio of these costs that; that means, how much more in comparison to a particular class how much more costly it is going to be misclassified the members of other class. So, we have the ratio of p p are able to understand the ratio or find out the ratio of these 2 numbers that would suffices to minimize this particular this particular quantity or expression.

Now, sometimes the samples that we use that we use for building our training partition might have one particular probabilities, but the new data on which we are going to apply our model might not have the same proportion of class same proportion of records belonging to different class members.

So, the in incorporation of prior probabilities can further improve the model. So, in such cases when we incorporate prior probability our average misclassification cost this particular formula will also change, but even in that case also the formula would change

and it will become C 0 multiplied by the probability value p 0 n 0 1 plus C 1 p 1 multiplied by n 1 0.

So, we can again take the we do not need to again we can take the prior probabilities even in this case and incorporate that in our model. So, we would not have to again we would not have to again find out the actual values and even in that case the constant term can be separated and we just have to know the ratio of these values.

So, that is why you would see that that is why you would see that many software's they would actually manage to discuss after and they would actually ask you to specify the ratio of misclassification ratio of cost misclassification costs belonging to different observation, belonging to different class observation, belong to different classes and they also incorporate the prior probabilities.

So, let us go back to our excel sheet and the same information is being conveyed over here the average misclassification cost of misclassifying a class 0 observation that we need to incorporate cost of misclassifying class one observation that we need to incorporate and you would see that if it is cost as q 0 q 1 and prior probabilities p 0 p 1 and we will have to change our sampling routine. So, earlier if we just do the simple random sampling and that might not work in case of prior probabilities because z if there is a rare case we will have to do over sampling for that particular case. So, for that we use stratified sampling.

So, in though in such cases we will have to minimize not just this this q 1 q 0 or will also have to minimize this probability of a particular record belong to class 0 and ratio of ratio of probability of class 0 membership and risk and probability of class 1 membership. So, depending on the case right we can minimize the particular quantity. So, we will stop our discussion here and will discuss further of asymmetric misclassification cost and some other concept under performance matrix.

Thank you.