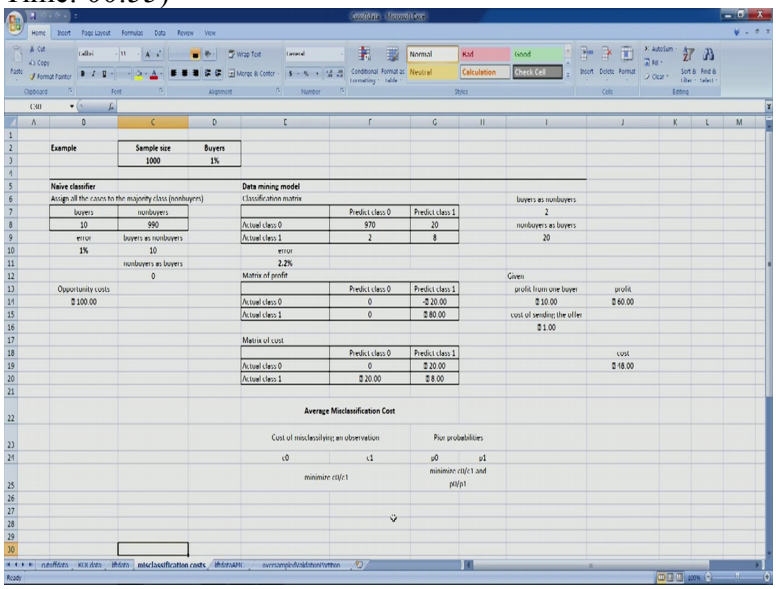
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture - 19
Performance Matrix-Part IV Assymetric Classification

Welcome to the course business analytics and data mining modeling using R. In the previous lecture we were discussing performance matrix. So, we were specifically we were discussing asymmetric misclassification cost. So, we will pick up our discussion from there. So, as we were talking about the asymmetric misclassification cost and how it can be important in a business context right. So, let us open the excel file so, that we are able to recall some of the thoughts that we had in the previous lecture.

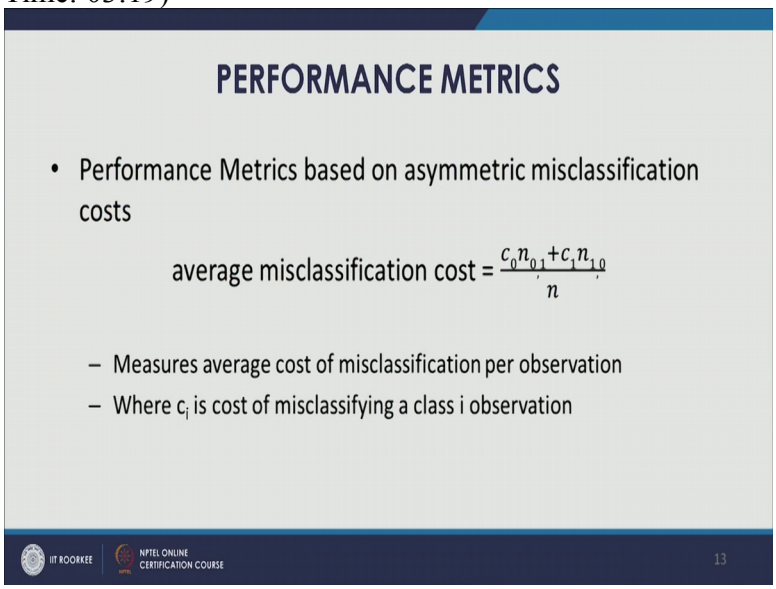
(Refer Slide Time: 00:55)



So, we talked about this particular example where the sample size is 1000 and we have 1 percent buyer, others being non buyer. We talked about what will happen in a new classified case, and we will have 1 percent of error if we classify all the records as non-buyers therefore, 1 percent bars would also be classified as non-buyers. And we will have 1 percent error we also have a hypothetical data mining model where the results would be in this form in this form of this classification matrix where we have this prediction that 970 members classified correctly and correctly as class 0 members and 8 members classified correctly as class 4 members others are misclassification errors.

So, that would lead us to 20 plus 2 that off diagonal elements and 2.2 percent errors. So, we talked about that this despite being a more misclassification error, higher

misclassification error, this particular data mining model could be preferable to us because, of the profits that would be involved the value that we can get from an customer selling a particular item to the customer.
So, that also we saw through this example where we had this matrix of for profit when the focus is on the profits, and then we also had this matrix of matrix of cost.
So, we have gone through this so, in matrix of profit we saw that through this is example exercise that 60 rupees profit we could make for buyer, and then we through a matrix of cost this is 60 rupees profit is from all the from the whole data mining exercise from all the buyers then we had matrix of cost from the cost perspective we had 42 8 dollars. So, our purpose is going to be when we take profit and cost into account the purpose is going to be maximization or profit or minimization of cost, but if we look at the improvement in our classification model. So, we would not see much in using these wage.
So, therefore, we talked about another performance matrix that was average misclassification cost.
(Refer Slide Time: 03:19)



So, this was the formula where we take so, this particular average misclassification cost as we talked about measures average cost of misclassification per observation, and this is the formula that we also discussed in the previous lecture right. So, we also talked about that this product we look at this particular formula the values n 0 1 and 1 0 and being can be considered constant and therefore, the minimization of this quantity that we require would essentially be based on the ratio of c 0 and c 1 right.
So, we would essentially be and it would be easier for us to estimate the ratio of c 0 and c 1 right cost of misclassifying a class 0 observation or class 1 observation that it would be easier for us to estimate that and therefore, essentially the software which

will try to minimize this particular expression average misclassification cost would essentially be doing on the basis of this particular ratio.

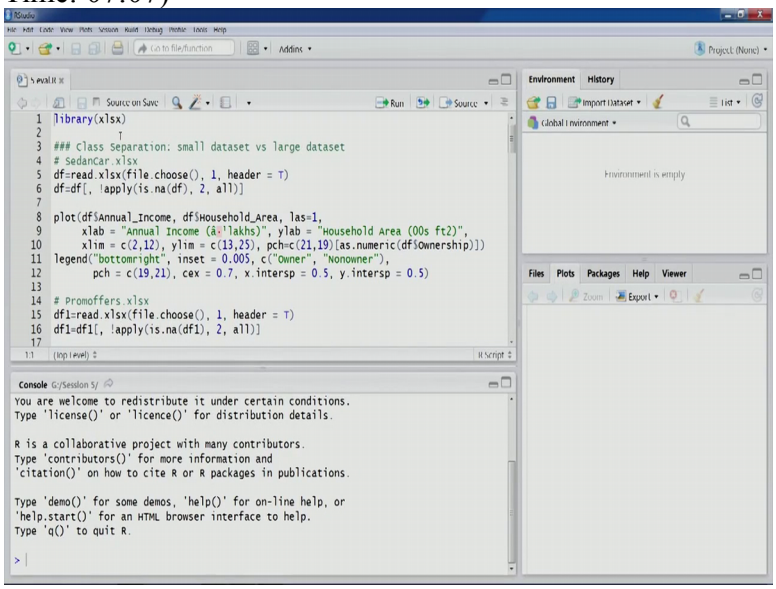
You also talked about that is why many commercial statistical software they will ask you to specify if there are any misclassification cost, you also talk about that some software will also ask you to specify if there are any prior probabilities. So, that was the discussion that we were having so, why we would be requiring prior probabilities in the in the first place. So, sometimes it might happen that the model we are building our model on a particular training partition, and the ratio of different classes might not remain same when we apply our model on new data. So, the ratio of class 0 members and to class 1 members are number of class 0 members and number of class 1 member that proportion, might not be same in in the new data all in the sample data.

So, if while we are building our model and the sample that we are going to use to build our model, and that sample might not have the proportion of class 0 members to class 1 members as would be there in the real data in the actual data that would also be used for testing, then we might incorporate that in our modeling process. So, how we incorporate that by specifying the prior probabilities so that is p 0 and p 1 p p 0 for class 0 members and p 1 for class 1 members. So, even in this case also this we would be just taking the ratio of these 2 prior probabilities would suffice us to this minimize the expression. So, we would essentially be minimizing p 0 divided by p 1 and c 0 divided by c 1. So, this is the expression that would essentially be minimized by the softwares instead of the full expression.

Now, we also talked about previously the importance of lift curve and how it can be used to understand to evaluate the effectiveness of a model.

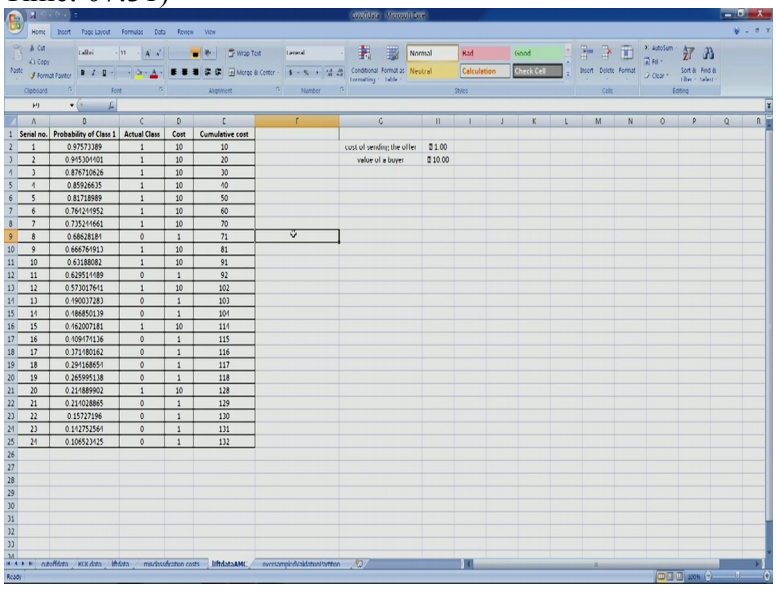
We saw that in comparison to a random selection case how the lift curve can tell us about the lift that we will get for a particular model, as the number of cases increase right. So, we saw that through this deciles chart also can we generate a lift curve while in court incorporating cost. So, let us do an exercise where we do this. So, let us open our studio.

(Refer Slide Time: 07:07)



So, as usual we will first load this particular library, data set that we are going to use it is particular again is in the same sheet or cutoff data dot xlsx. So, now, let us have a look at the data as well.

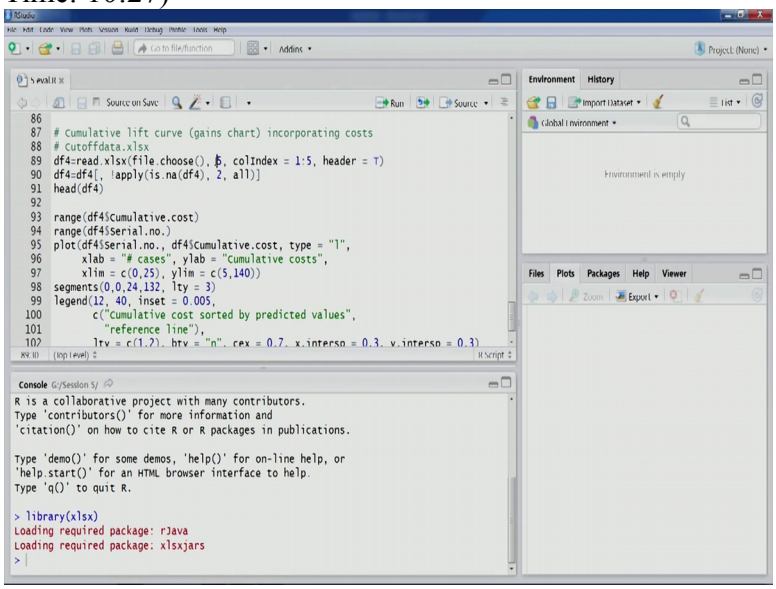
(Refer Slide Time: 07:31)



So, the data set that we are going to use is this one. So, you would see that 5 variables are there, but if you look at these are based on the results. So, we have serial number that that representing the index of particular observation, and the data has been sorted based on the estimated probabilities of class membership specifically class 1.

So, based on the estimated probability of class 1 the data has been sorted, we you can also see then that actual class in the third column has also been mentioned. So, that particular observation belonging to the actual class of that particular observation, and

the probability of class 1 is given there. So, point 5 of anything is if any probability i value being more than point 5. So, that class would be predicted as the 1 otherwise 0.
So, now, in this particular data set you can also see that we are trying to incorporate the cost, the concept misclassification cost concept that we discussed. So, you can see your cost of sending the offer we have specified as 1 rupee and value of a buyer thirty have we specified has 10 rupee. So, now, at this point I would also like to tell you that this is a small data set. So, the plot that we are going to generate would be slightly different in comparison to if we had full data set, bigger data set.
So, from this example you would also know that there are 12 ones and 12 class 0 members, and 12 class 1 members in this particular small data set. Now the depending on the actual class; however, specified you can look at the excel formula that is there depending on the actual class I have a specified the cost. So, that is if the actual class is 1 then value of a buyer that is 10 rupees that would be taken up if it is 0, then the cost of sending the offer that would be used. Once this value is there then we can also find out the cumulative cost you can see the cumulative cost here. So, for to compute the cumulative cost for each observation for the first observation it is going to be the same as cost for the next observation, we are going to add the value of previous result as well.
So, in this fashion will continue and up to the last observation. So, will have this so, essentially in the lift curve we would be plotting this number cumulative cost number with respect to the index of the observation. So, let us open our studio. So, you would see there is slight change in the function read dot xlxs that we are going to use here.
(Refer Slide Time: 10:27)



You would see that I have taken this sheet index as 5 this is 5th work sheet, and you would see also that column index and other argument have used. So, we just want to

we just want to import the data of first 5 columns. So, that is starting from 1 to 5 only these so, we are interested in putting only this data that us right other details this particular detail cost of sending the operand value of y this we do not want to include in our imported data set that execute this line.

And we will get the appropriate data set into our environment. So, let us look at the data that we are going to use for this slept curve generation. So, we have 5 columns here serial number that is representing the index of the observation, and then a probability of class 1 so, this particular data set has been sorted from higher probability to lower probability the estimated probabilities by the model. Small data set we have 24 observation, and we have 12 observation belonging to class 1 and 12 observations belonging to class c class 0, the we have third column is also mentioning the actual class of particular observation, and then we have cost which is determined by.

So, we have some numbers here you can see that cost of sending the offer that we have specified is 1 rupee and value of a particular buyer is 10 rupees.

So, therefore, if the offer is accepted by buyers. So, the net value that the impact that we that the organization get gets from a buyer is rupees 10, and if that offer is not accepted by the buyer then it would cost them 1 rupee in the cent and the cost this is actually the cost of sending the offer.

So, you can look at the excel formula has been appropriately specified, where if the actual class is 1 right in that case the value of buyer 10 is mentioned this is net value. So, that is mentioned if the class is if the class is 0 and you can see then in the other in that case this particular value minus of this value is actually mentioned. So, in this fashion for every observation which has been sorted by the estimated probability of belonging to class 1 we can mention the cost.

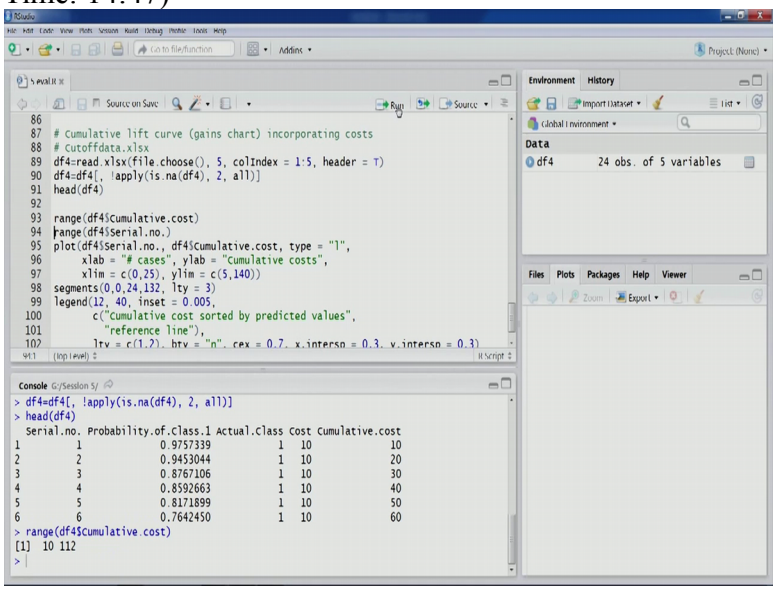
So this will this this per these particular numbers in this particular column cost they are including the net value from a buyer or and the cost of sending the offer as well. Now cumulative cost has been computed using this particular cost column. So, in this case first observation the cumulative cost is the same as the cost for the first buyer, then the for the second buyer we are adding up the cost that incurred in the first buyer.

So, in this fashion for previous cost is being accumulated and we get this variable cumulative cost, so let us import this particular data in the r environment so, you can see some changes in the function v dot xlxs that I am calling here you can see that work sheet that we are calling is 5 you can see in the excel file. So, this is work sheet number 5, but you can see this work sheet number 5 Otherwise you can also import

the data from this worksheet by using the name of this work sheet with which we have provided here as well.

So, right now we are going to use the index then you can also see the columns that we want to import. So, we want to import column number 1 to 5. So, we are using column index argument here you can see we are interested in for the column 1 to 5 this program formation we do not need for plotting lift curve. So, let us import this particular data set you can see 24 observation of 5 variables let us remove if there are any columns.

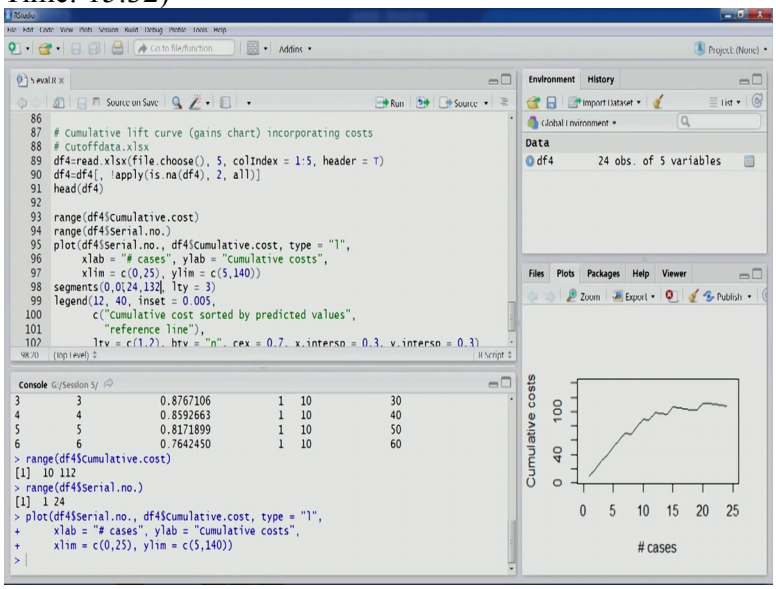
(Refer Slide Time: 14:47)



Let us look at the first 6 observations you can look at the first 6 observations cumulative cost is there. Now we are going to we are going to generate a scatter plot between these 2 variables the serial number, and the cumulative cost data cost.

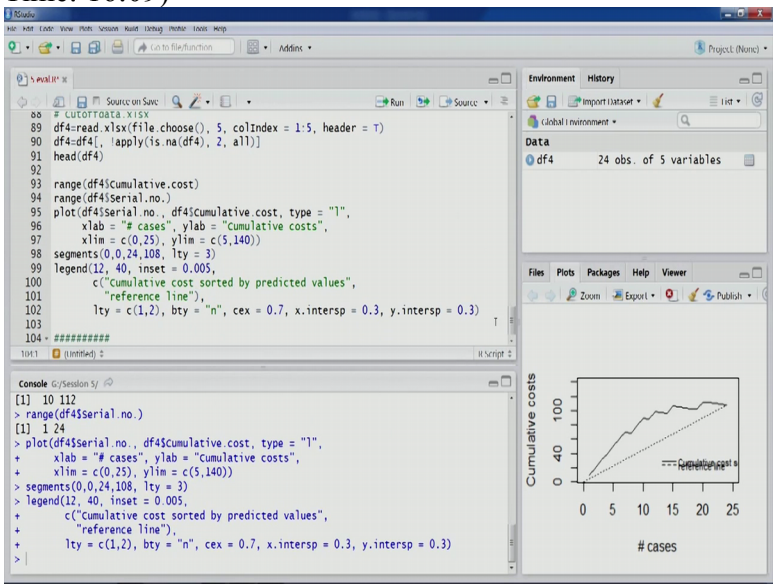
So, let us look at the range of these 2 variables can see that range is 10 and 112 for cumulative cost and for serial number as we know that there are 24 observations. So, you can see that these limits have been appropriately specified from 0 to 25 and also from 5 to 140. So, all the values would be covered in this particular range.

(Refer Slide Time: 15:32)



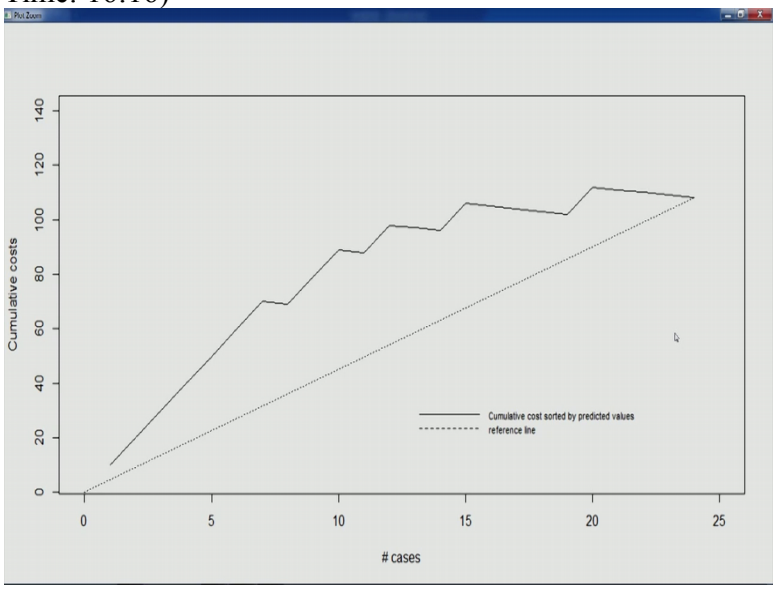
So, let us create the scatter plot you can see this plot, let us zoom in let us also generate the reference line. So, in this case reference line is conducting the initial point that is in this case for reference line we have taken a 0 0, and the last point that is twenty fourth observation and the corresponding value. So, in this case the corresponding value is let us confirm using excel file it is 108.

(Refer Slide Time: 16:09)



Let us generate this line as well and legend. Now let us look at the zoom plot now we look at the if we look at this particular plot.

(Refer Slide Time: 16:16)

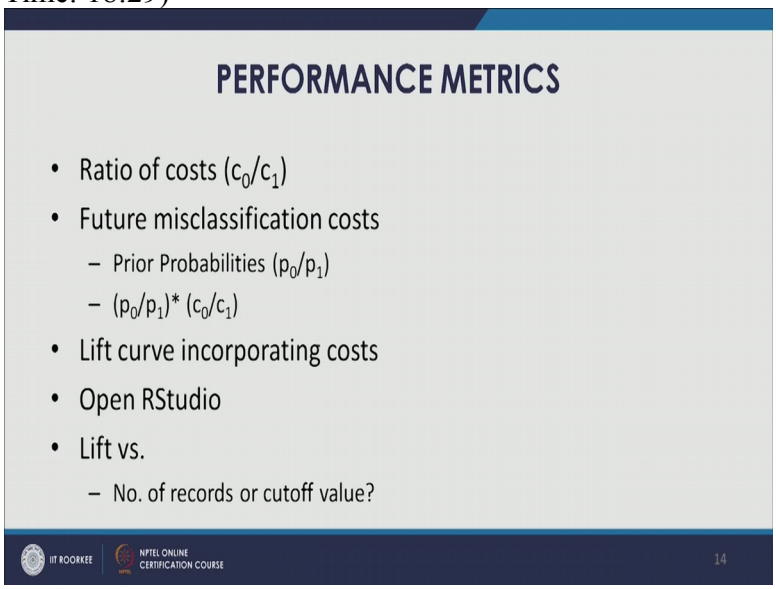


So, this is essentially plotting the accumulative cost in the y axis and the number of cases in the x axis. So, a dotted line being the reference line which is connecting the 0 0 to the last point you can see the cumulative cost is continuously increasing as we as the number of cases increased from as we move along the x axis from left to right, but you would see at the end of it there is slight dip. The same thing is reflected in the excel sheet that we see you can see here, that the maximum value that we have is 112 and after that you would see there is a dip and the last value is 1 0 8.

If we had a much larger data set a many more observation this particular dip would have been even more to the extent that this line might come down, the slope of this line might come down much more and it can even be negative for example, if we consider case where we have let us say 20000 observations. And they are the high most probable ones would be selected in the initial part of the plot and the buyers the non-buyers would come later and thereby they will decrease the last value of the last point, and the whole plot might go below the x axis some part of the plot might go you know go below the x axis and therefore, the slope of this reference line might even become negative. So, it can come down

So, you can see the optimal point is this particular point where the value is maximum which is 112 to this point this is twentieth observation so, this is where we are getting the optimum value. So, let us go back to our discussion. So, that was the lift curve. So, lift curve can actually be used to find out the most probable ones the rank ordering.

(Refer Slide Time: 18:29)

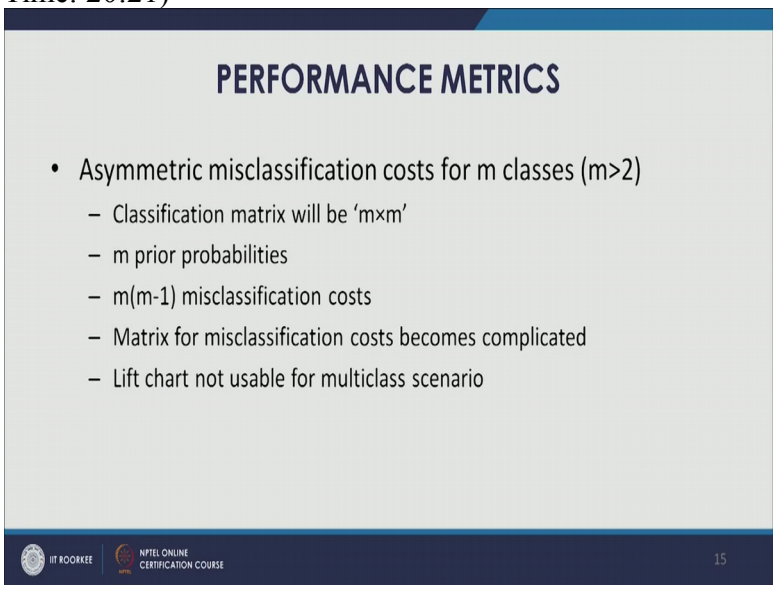


And how many probable ones can actually be sent the offers first, and what is the optimal point how many buyers would actually be sending the offer. So, around the business context that would be more desirable. Now lift curve that we just plotted. So, there would be 2 scenarios so, 1 is the plot that we just generated was lift versus number of records. So, the goal was to identify the goal was to see the cumulative cost that could be there the optimal point, that where the optimal point could be and also to identify the more likely it is buyers which are more likely to accept the offer.

Now, we can also generate a lift plot lift curve versus cutoff values. So, if we are interested in finding out the cutoff value where the model is going to perform better. So, that can also be found that can also be done. So, we can also have so, depending on the goal whether we are looking to find the suitable cutoff value or whether we are looking to identify the check the effectiveness of the model, and the number of records most probable records. So, depending on that we can plot the lift curve accordingly.

Now, whatever discussion that we had about the asymmetric misclassification , and other concept before that was mainly applicable in 2 class scenario class c 1, and class 1 where class 1 was ever class of interest right can this all these concept can they be extended to m class scenario.

(Refer Slide Time: 20:21)



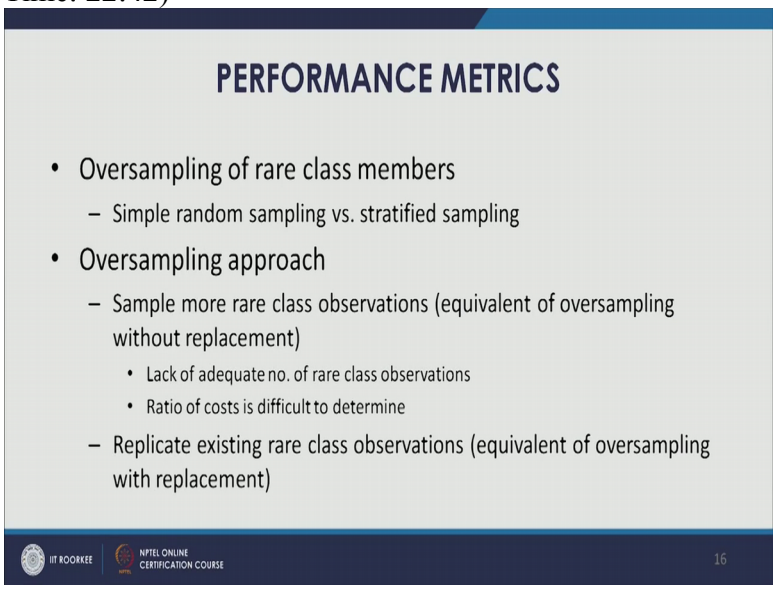
So, if we extend asymmetric miss classification cost for m class scenario, that is m greater than 2 we will have a classification matrix m class 1 m rows and m columns. So, this is going to be a much bigger classification matrix therefore, the kind of computations and discussions that we had they would become even more complicated. So, we will have to deal with m prior probabilities if suppose that sampling is distorted, and the real data in the original data the proportion is different then we will have to incorporate m prior probabilities

Similarly, we look at the misclassification cost there could be m times m minus 1 misclassification cost. So, the understanding different misclassification cost that could be there that would also become slightly more complex. Lift chart that we have generated for 2 class case that would also be that can also not be done in a multiclass scenario until, and unless we identify our earlier class the class of interest and other classes we combine and reach to the 2 class scenario only then it could be used.

So, otherwise the discussion that we discuss on related to probability and misclassification cost they can be easily extended, but when we want to execute the things are going to be when we want to do an actual implementation of this it would be more complicated. Next point that we want to discuss is over sampling of rare class members. So, sometimes it might be the case that the class of interest might be having very few class members. So, it might so if we are want to create a model with very few class members the model might not be really useful. So, in this case we would be required to do over sampling of rare class members.

So, that will bring us to our discussion on simple random sampling versus stratified sampling. So, when we do simple random sampling when we have a rare class of interest with very few class members belonging to that particular class then.

(Refer Slide Time: 22:42)



Simple random sampling might not give us the good enough partition to build models right, and your training partition if it is randomly drawn we might not get enough number of cases belonging to class of interest, and that can impact our model and then later on it is up implementation on new data. So, in such situation we will have to do over sampling as I said and stratified sampling is generally used for this kind of tasks. So, generally stratified sampling is used to perform over sampling especially, if there are such groups are present where 1 group is dominating the and other group is very few members of other groups are present.

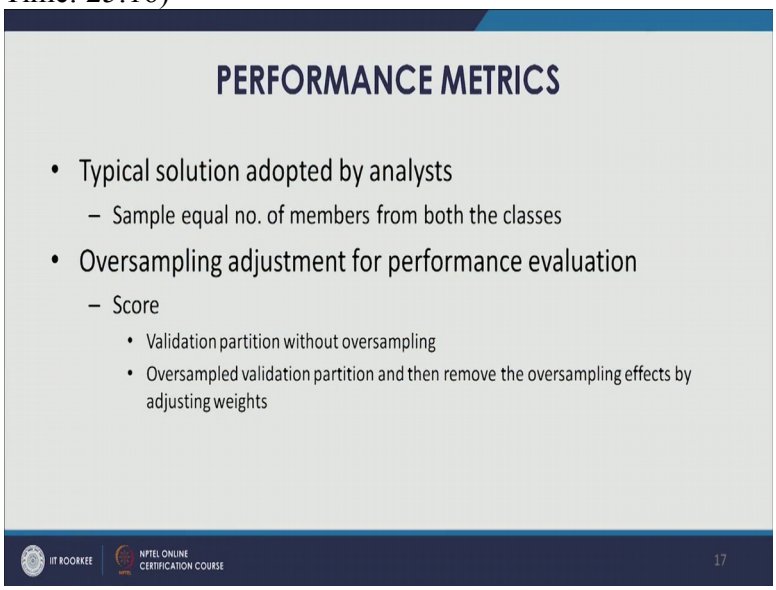
So, what could be the different over sampling approach so, one way could be sample more rare class observations. So, this is like equivalent of over sampling without replacement. So, the sample that the data set that we have we can sample more of the class 1 members the class of interest right and so, the problems that can be faced in this. So, this is more desirable approach, but there could be some practical problems that we might have to encounter. For example, lack of adequate number of rare class observation what if in the data set itself the number of rare class observation or so, few that even this you know sampling of more real class observation might not feel a meaningful you know sample and therefore, meaningful modeling.

So, those practical problems we might have to encounter. Now the ratio of cost that will that is also difficult to determine, when we are faced in this this situation where very few you know members are present this ratio of cost, this is also difficult to

determine. Another approach could be replicated existing rare class observation. So, this is equivalent of over sampling with replacement. So, some of the observations that are present in our data set we can replica we can have copy replicates the same observation, and then use that for the modeling

So, this is another approach what the analysts generally do typical solution that is adopted by analysis sample equal number of members from both the classes.

(Refer Slide Time: 25:16)

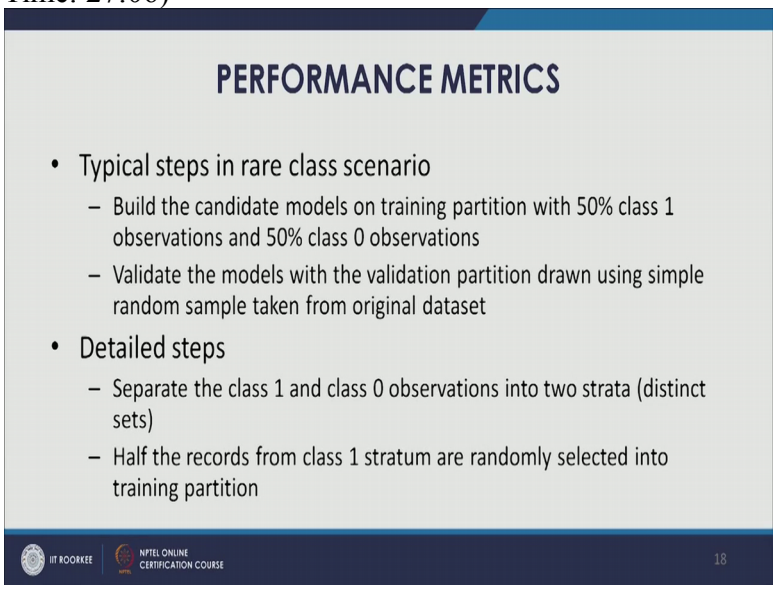


So, what they generally do is they take 50 percent of the members from class 1, and 50 percent same number of not 50 percent same number of may same number of members from both the classes, and then use that for their analysis. Now if you have done your modeling using an oversample training partition right even if it is only for the rare class, then the performance evaluation you have to adjust for that over sampling.

So, when we go about is when we use our model that has been developed on over sample partition will have to score this particular model right so, validation partition validation the one approach could be scored the validation partition without over sampling. So, that validation partition that does not have the over sampled cases over to sample the cost so that can be used.

So, that is the easier and direct and straightforward approach. So, build your model on over sample training partition, and evaluate your model using a validation partition the from taken from the original data set. Now another approach could be use the over sampled validation partition, and then remove the over sampling effect by adjusting weights. So, these 2 approaches could be there. So, the steps typical steps that are taken in rare class scenario. So, build the candidate models on training person with 50 percent class 1 observations and 50 percent class 0 observations.

(Refer Slide Time: 27:06)



We take equal numbers and validate the models, with the validation partition drawn using simple random sample taken from original data set detailed steps. So, we will stop here, and detail is step of this particular procedure we will discuss in the next lecture.

Thank you.