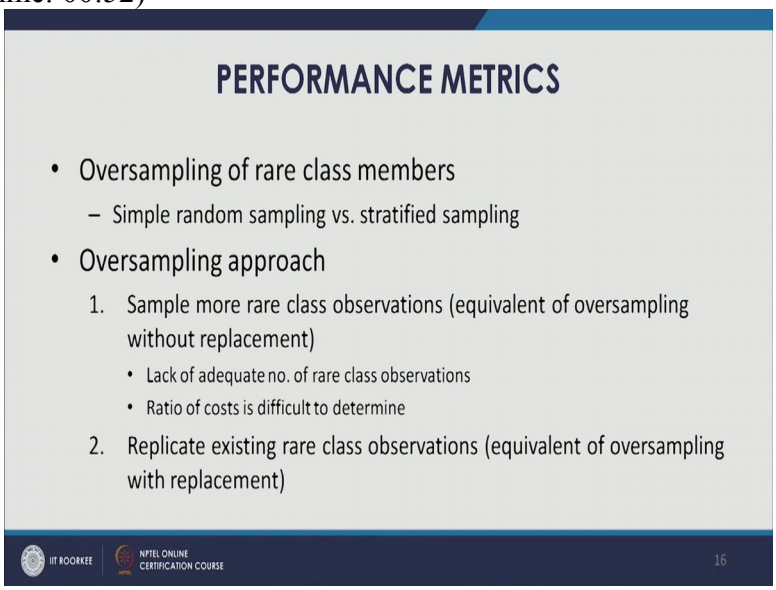
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture - 20
Performance Matrix-Part V Over Sampling Scenario

Welcome to the course business analytics and data mining modeling using r. So, in the previous lecture we were discussing performance matrix, and specifically at the last part of the lecture we were discussing over sampling approach especially in the scenario, where you are dealing with rare class members where, the class of interest members belonging to the class of interest they are very few in the sample. So, we talked about different things related to this particular approach.

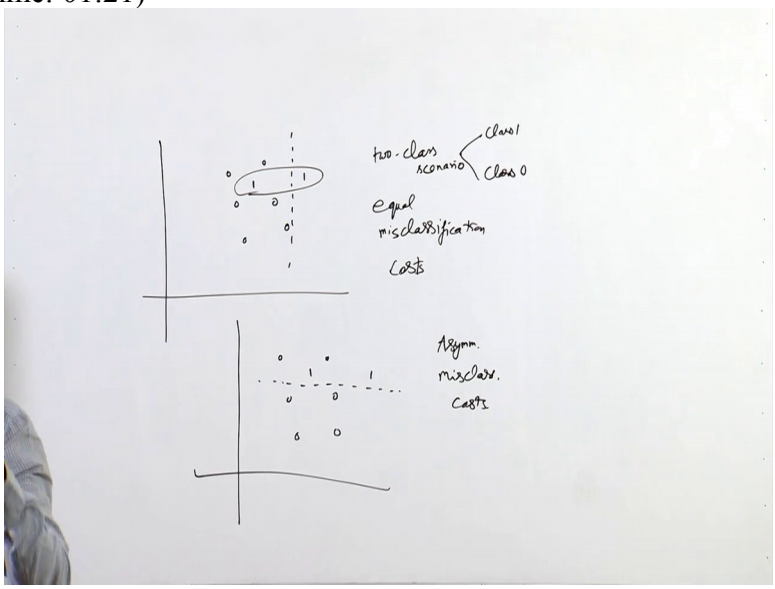
(Refer Slide Time: 00:52)



We talked about that there could be 2 over sampling approach, 1 being that we can sample more rare class observations, and the second being that replicate existing rare class observations.

So, we focused more on sample more rare class observation which is equivalent of over sampling without replacement. Let us also understand few more situation through using some graphs.

(Refer Slide Time: 01:21)



For example when we are dealing with the when we are dealing with 2 class scenarios where the classification cost are equal. So, we are dealing with equal misclassification cost, the scenario where we are dealing with equal miss classification cost and we have 2 class scenario as we discussed us in the previous lecture and these 2 classes being class 1 and class 0.

So, if there are more records belonging to class 0 being represented by 0 itself here in this graph, and there are few records belonging to a class 1 right, and we are in the scenario where we have equal misclassification cost. So, a 1 particular model could be this one so, this line will separate the records will create homogeneous partitions and also minimize the misclassification error right. This is homogeneous all they record there is just 1 record there belong to the same class, and here most of the records belong to the class 0 just 1 misclassification error.

So, this is the case when we are dealing with the equal misclassification cost, and this is how it is going to work out. The another scenario could would be when we have asymmetric misclassification cost the same scenario 2 class scenario class 1 and class 0 in such a situation we might be interested in as we discussed, in the previous lecture we might be interested in identifying more of the class 1 observations. Even if it comes at the expense of misclassifying more of class 0 observation.

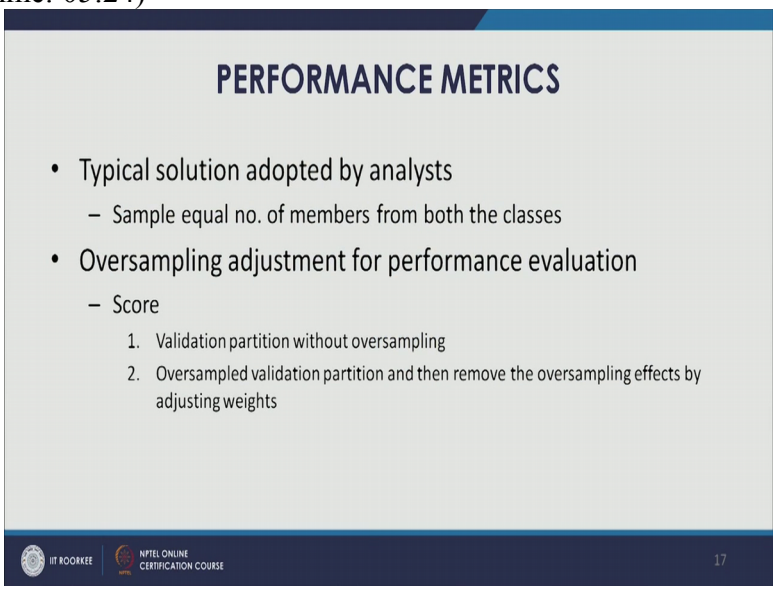
So, if we try to plot the same points here. So, our model could be represented by this particular separator line right this particular separator, now you would see that in the upper partition we have all the observation belonging to the class of interest that is class 1, and in the lower half we have all the observation this is homogeneous belonging to class 0. So, this is more desirable when we have asymmetric misclassification cost that

meaning, that there is 1 specific class of interest and we would like to identify more of that class, even if it comes at the expense of misclassifying some of the other observations.

Now, when we talked about this over sampling scenario that these rare class members they could be very few in the sample, what we can do about that, so we talked about that over sampling to over sampling approaches could be used, sample more rare class observations, that is equivalent of over sampling without replacement then another one being replicate existing rare class observation. So, we further talked about what is the typical scenario that is followed by analyst. So, they generally sample equal number of respondents from both the classes right so, that is the typical approach.

So, we talked about when we follow this approach irrespective of the over sampling approach that we follow.

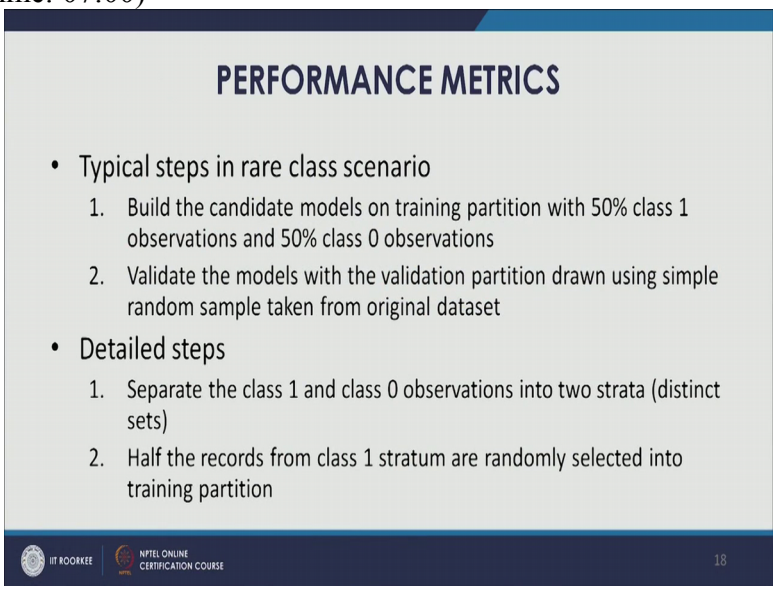
(Refer Slide Time: 05:24)



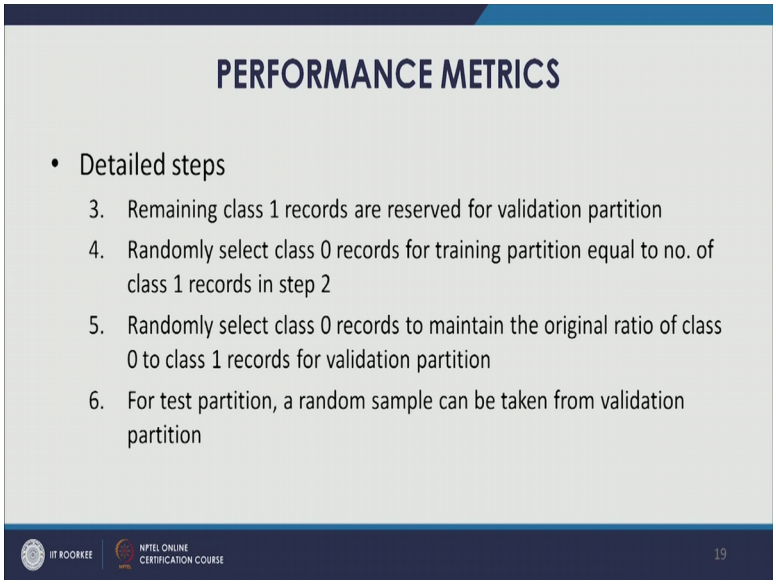
That over sampling adjustment will have to perform to for our performance valuation, and to evaluate the performance of the model right, we talked about 2 particular scoring methods for validation partition. So, one first one was to build your model on over sample training partition, and then test it on a regular validation partition. So, that is indicated by the first one validation partition without over sampling that is regular or the original validation partition. Now second one being over sample validation partition, and then remove the over sampling effects by adjusting weights.

Sometimes the number of a rare class cases could be class of interest cases could be so few that even that validation partition without over sampling might not remain practical might not remain useful, in such situations we might have no choice, but to over sample even the validation partition. So, we will be building our model on over sampled training partition as well as we would be evaluating our model on over sampled variation

partition in those situations. Now we also talked about few typical steps in rare class scenario that are generally taken. So, as we discussed that equal number of observation from both the classes the same thing can be said that training partition with 50 percent class 1 observation and 50 percent class 0 observation so, equal number of participation.
(Refer Slide Time: 07:00)



Now validate the models this is the main this is the approach 1 that we actually would like to follow that validate, the model with validation partition drawn using simple random sampling taken from the original data set. So, the regular data set. So, we would like to have validation partition taken from the regular data set. Now to clarify these steps a bit more to have the detailed step of this particular process, we have listed them out all these steps the first one being separate the class 1 and class 0 observations into strata. So, we talked about in the previous lecture that a stratified sampling is generally used for over sampling. So, the first very first step itself is talking about the same.
So, we can now whatever the number of observation that we have in our sample, we can separate the class 1 observation from class 0 observations, and we can create 2 strata or 2 distinct sets of those classes. Now because we would be requiring training and validation partition at least, therefore, it would be advisable to use half of the records from class 1, we can select half up those class 1 records randomly and then put them into training partitions, the remaining half of class 1 records they can be reserved for validation partition or even for test partition we will see as in the next steps.
(Refer Slide Time: 08:56)



So, half the records from class 1 observation class 1 stratum can be randomly selected. The remaining class 1 records are reserved for validation partition. Now next step would be randomly select class 0 records for training partition equal to the number of class 1 records that we did in step 2. So, that we are able to maintain that 50 percent records coming from each of the classes right for both from the classes we would like to have these equal number of records. So, what we did in step 2 the same number of records we can also randomly select from class 0 stratum.

So, the next step would be randomly select class 0 we got so, till step number 4. So, we have been able to create our training partition. So, we can do our modeling, so we can build our model then later on when we will require to evaluate that performance of that particular model we would require validation partition.

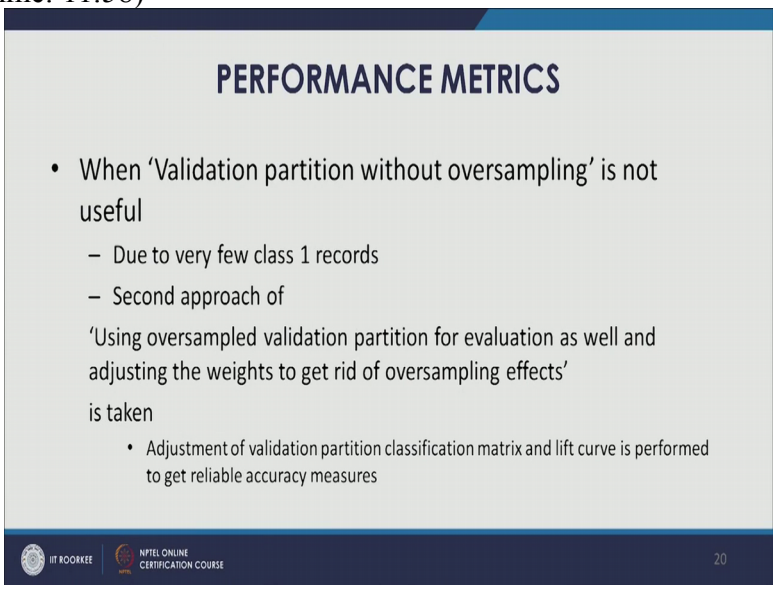
So, as step number 5 onwards they deal with creation of validation partition. So, you would see step number 5 is randomly select class 0 records to maintain the original ratio of class 0 to class 1 records for validation partition, because we said that our first approach would be to test the performance of this being the direct approach also that build your model on over sample training partition and test it on regular validation partition.

So, to build the regular evaluation partition we would like to randomly select class 0 records and the half of class 1 records we already have so, we would like to maintain the original ratio that was there in the regular dataset. So, in this fashion we can create a validation partition. Now this could be used for performance simulation, but if we want a further you know test partition because sometimes the validation partition could also be used for fine tuning the model and therefore, it also becomes part of the building process of the model.

And therefore, test partition is very much required to evaluate the true performance of the selected finally, selected model right. So, therefore, if you want test partition also so, the validation partition that we might we have created during step 5. So, from that validation partition itself we can randomly draw a sample for test partition. So, in this fashion we can perform our modeling we can over sample, if there are you know class of interest there are very few cases we can do over sampling for the training data set, and we can prepare our validation partition accordingly. So, that it is i as per the proportions in regular data sets.

And therefore, we can go ahead and evaluate the performance. Now there as we discussed that there might be situation where validation partition without over sampling the first one that we just talked about the detailed steps.

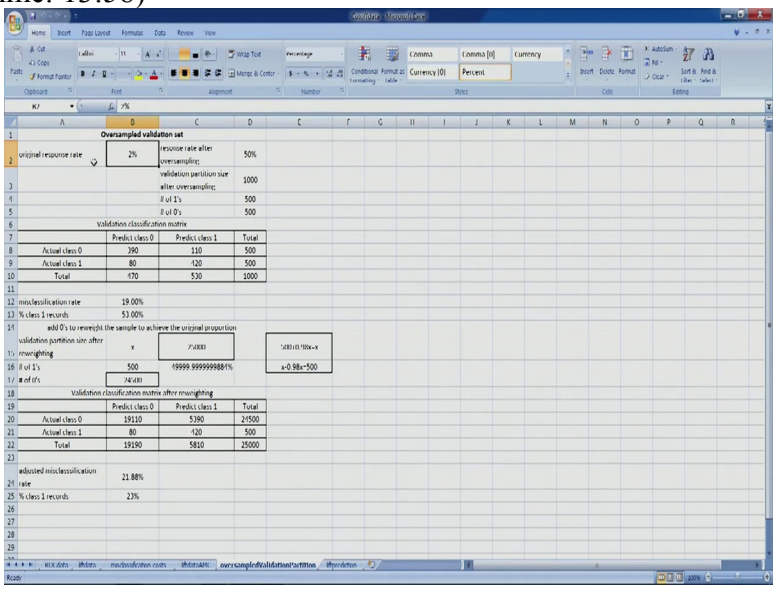
(Refer Slide Time: 11:58)



We talked about that might not remain practical that might not be useful reason being due to very few class 1 records that you know some few records that, half of the records for training partition and reserving half of the regards for a validation partition. Even that might not be feasible or practical and therefore, the modeling might the modeling might become a bit more complicate.

So, we can we have to follow the second approach was to over sample even the validation partition for evaluation as well. So, this validation is going to be used typically for evaluation. So, we are going to use the over sample validation partition, and later on we will have to adjust the weights to get rid of over sampling effects. So, we apply our model that has been built on over sample training partition, then we test the performance of that model on the over sample validation partition then we readjust the weights that we get there and so, that we can remove the we can get rid of the over sampling effects.

So, now because this is main idea is about evaluation of the model. So, adjustment that we require is on a validation partition classification matrix that we get when we test our model on validation over sample validation partition.
So, that adjustment is required there and lift curve can also be adjusted accordingly. Now what we will do we will do an exercise in excel for this whatever we have discussed till now, let us go through an exercise. So, this is the scenario we have over sample validation set.
(Refer Slide Time: 13:58)



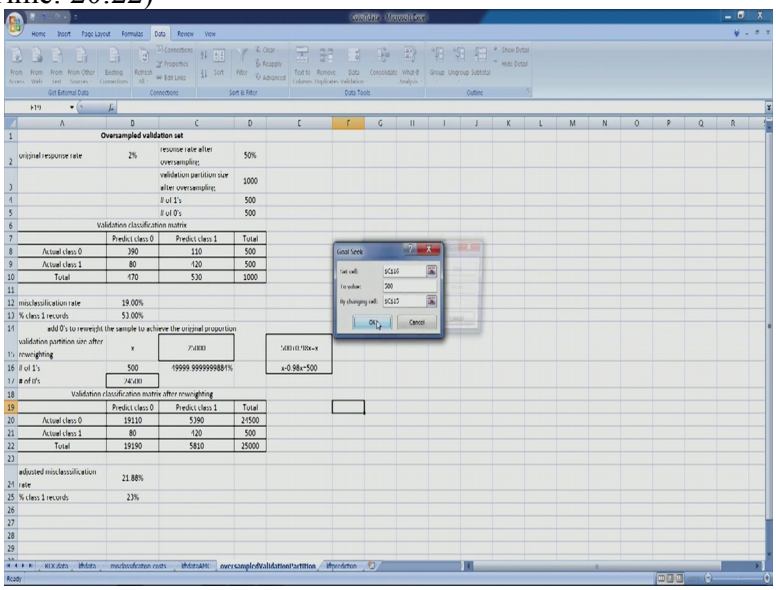
Now, the assumption is that the sample that we have the original response rate is 2 percent; that means, the response variable the records belonging to that class of interest, they are just 2 percent and the other records belong to the other class 0. So, recorded belonging to class 1 they are just 2 percent of the sample, and the 98 percent of the records in the sample they belong to class 0.
So, when we do over sampling we try to increase this particular ratio this particular portion, and we make it 50 percent. So, we over sample in such a manner that now the response rate it increases from 2 percent to 50 percent. So, now, in this over sampled data set that we have will have 50 percent records belonging to class 1 observation class of interest and the remaining 50 percent belonging to the class 0 observation, same you can see that validation partition size after over sampling if it is 1000, then the number of ones in this particular 1000 records are going to be 500 and number of 0s are going to be 500.
Now let us say we build our model on it on the oversampled training partition, and then later on we applied that that particular model on over sampled validation partition. So, as a result of that that validation exercise we got this classification matrix this validation classification matrix. So, you would see so, because there were more than usual response

rate because of the over sampling you can see the results, this could be 1 example of this classification matrix.
So, you can see 390 for a class 0 members classified as class 0 than we have 420 class 1 members classified as class 1 members. So, in total you have 500 class 0 500 class 1 total 1000 records in the sample right. So, if you want to compute the classification laid this in this fashion as we have discussed before that off diagonal elements that is 80 and 110, and then it would be divided by the total number of records.

So, that will give us the misclassification rate that is a nineteen percent. So, when we do over sampling this is the misclassification rate that we get, but this is on the over sample validation partition. So, this is going to be slightly less than what could have been there in the regular dataset scenario. So, if you look at the number of percentage of class 1 records so, you can compute this also right so, this comes out to be 53 percent.
So, 53 percent records have been classified by the model as belonging to class 1. Now to access the to evaluate the true performance of the model that we need to adjust the weights so, there are 2 ways either we remove you know ones so, that we are able to get maintain the original proportion or we can add 0s. So, that we are able to again and get back to the original proportion. Now the typical strategy that generally we follow is adding 0s to reweight the sample to achieve the original proportion, now how that can be done you can see that let us say validation partition size after reweighting reweighing is x right.
So, we are going to use some of the utilities that are available in excel is specifically goal c this is an easy equation that we can solve manually as well, but because we are using excel so, we like to use some of the utility that are available there. So, if we are if you want to add extra 0 so, that we are able to reach the we are able to reach the original proportion so, we need to find out. So, this new this new sample size would be much bigger. So, let us so this is going to be the equation that will give us the new sample size. So, earlier 1 was 1000 now we want new sample size which will have the original proportion of different class members.
So, 500 if there are 500 and ones and we are not going to change the this particular figure. So, we would like to add 0s so, 500 representing the 2 percent so original response rate was 2 percent. So, 500 value of number of ones is representing now the 2 percent of the response rate. So, therefore, we need 98 percent of 0s here. So, we will this is how we can write this equation 500 plus 98 percent 0.
So, therefore, 0.98 into x x is the total number of records that is in the new sample reweighing and that has to be equal to the total number of records 500 plus 0.98 x equal

to x. So, if we solve for the value of x. So, we will get the new sample size after doing a weight adjustment. The same equation can also be written in this fashion x minus 0.98 x is equal to 500 this being more this being suitable for us to be able to use goal seek function in excel. So, if you want to use goal seek function which we have already done for example, we want to solve for x. So, this is the x and this is the cell that we have reserved for the value.

(Refer Slide Time: 20:22)



So, let us find out this. So, you go into the data tab and then you would see this what if analysis there and then there you would see that goal seek is there. So, the set cell there we need to specify the cell where this particular formula is there. So, formula what we have written in this particular cell right, and then the value that we want to target that is 500 right, you can look at the new equation so, this is what the equation that we are trying to target.

Now the formula has been written there so, we will have a look at the formula the formula is actually this representing this particular expression x minus 0.98 x we will just see what we have formula we have written in the particular cell c 16 right. So, the value that we are targeting is 500 that is right hand side of this equation. And we are changing the cell the this one c 15, which is representing the x.

So, we want to change this particular cell and if you just run through, we will get these values which are already regard they are there because I had ran this particular goal seeker utility before. Now look if you want interested in the formula this is the formula that we have written there. So, this particular formula is representing this equation x minus 0.98 x that is going to be used by the old c function as we just saw. So, you can see if c 15 this is the value c 15 is this x right so, c 15 representing the x minus 0.98.

So, that this is how we are computing this 0.98 1 minus b 2, b 2 is our original response rate. So, 1 minus b 2 it is in percentage will excel will take care of this percentage notation, and it will appropriately convert it into 0.2 and therefore, 1 minus 0.2 will become 0.98. So, we will get that value then again this is again 0.98 x.

So, again x being c 15; this is the formula that we have written there. Now as we saw that we can learn the goal seek function I will get this value 25000 here so, this is the 25000 is the size of a validation partition after reweighing, as you can see validation partition side after reweighing is 25000. Now 500 ones were there so, the remaining number of 0s would be can be very easily computed using this formula or manually as well for this particular case.

So, 24500 0s are to be there, now once we have once we are done with this particular calculation. Now we can adjust our matrix validation classification matrix using this particular information the new sample size, and the number of funds and the number of 0s. So, here as you can see in this classification matrix, now we can fix these values here you can see v 17 this particular value number of 0s, this has been fixed using this number then the this has also been fixed using this number of 1s total is also fixed using this number c 15, and then we are now we need to adjust the values that are there in the in this 2 cell the cells for class 0 members.

You can see that class 1 members they are unchanged right. So, because they were 500 we do not want to make any change there. So, these value remain unchanged as you know same as the previous matrix, another 2 values that we need to find out is these 2 values right. So, the how we can do this is we can keep the ratio that was there in this particular matrix that we got from the model and so, we need to maintain this ratio 390 to 1 390 and 110 for the total number of observation of 500. So, this ratio we need to maintain 390 divided by 500 to 110 divided by 500. So, this ratio for class 0 and class 1 we can maintain and we can compute this new number.

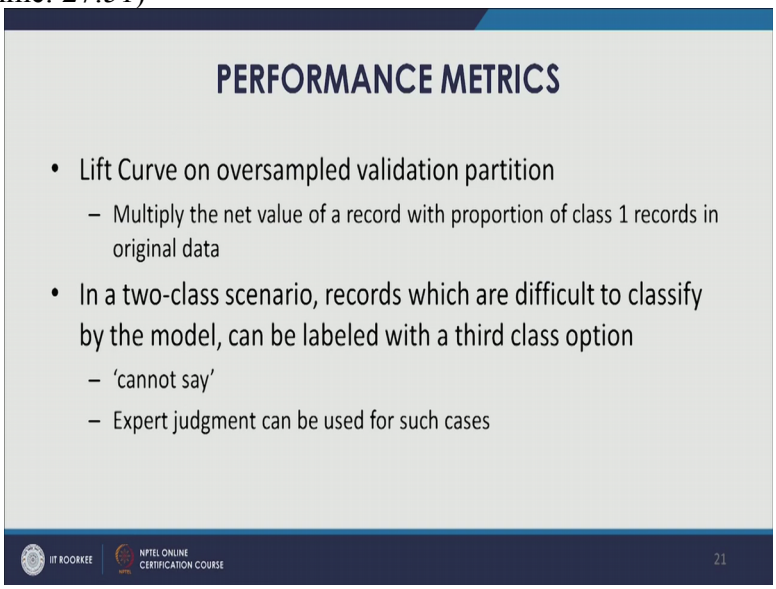
Total we already have so, in this fashion you can compute you can see this that 390 divided by a 500 this is the ratio, and total number of observations are 24500. So, we will get the new number of 0s in that particular cell. Similarly this one can also be computed to this particular cell value you can see 1 once again hundred and 10 divided by 500 as represented in c 8 divided by d 8, and then total value being 245, 1500. So, the number has been appropriately calculated by actual, now we can also get these values predicted class 0s number predicted class 0, and number predicted class 1 records using this particular some function. So, in this fashion we will get the new validation classification matrix after reweighing.

Now, once the weight adjustment has been done, we can use this particular new matrix to compute adjusted misclassifications a. So, now, following the same procedure as we having using before we can look at the off diagonal values to find out the error, and now we get the new number that is 21.88 percent. Now the earlier misclassification rate was 19 percent which was slightly lower than this number 21.88 percent so. So, now, once we are able to remove the over sampling effect you can see that miss classification has laid has jumped to almost 3 percentage point right, and if we look at the percentage of class 1 records right. Now they have they have come down. So, earlier we had 53 percent in the over sampled partition validation partition.

Now, the new numbers as you can see this is how you can compute c 22 divided by d 22 these 2 numbers right. So, you can see the new percentage is 23 percent of the records; they have been classified as class 1. So, let us go back to our slides. So, in this fashion what we were talking about that if we had to use the over sample validation partition, then how do we go about evaluating the performance of our model. So, we will have to adjust the weights. So, that we get the new validation matrix validation partition this classification matrix, and then that can be used to compute the misclassification error.

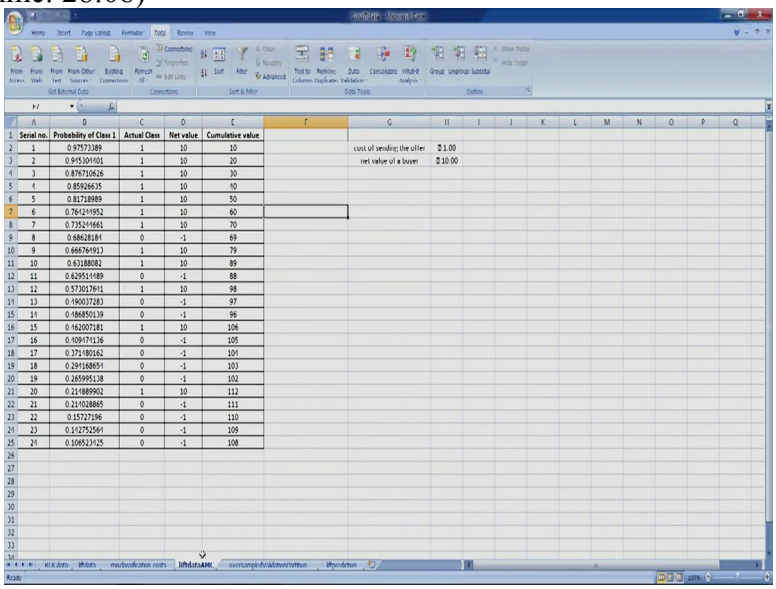
Now, the lift curve can also be appropriately adjusted so that we can again compare the find out the efficiency of the model for the over sample validation partition case. So, you can see the same thing is being discussed here, lift curve on over sample validation partitions.

(Refer Slide Time: 27:51)



So, how to create that so the steps that we have been following earlier right the lift curve that we had created before, the same steps can be followed here right lets go back to our excel file once again and let us look at what when we created a lift curve.

(Refer Slide Time: 28:08)



Serial no.	Probability of Class 1	Actual Class	Net value	Cumulative value
1	0.975773389	1	10	10
2	0.945304101	1	10	20
3	0.876710626	1	10	30
4	0.85926635	1	10	40
5	0.81718989	1	10	50
6	0.761241952	1	10	60
7	0.735241661	1	10	70
8	0.68628184	0	-1	69
9	0.666766913	1	10	79
10	0.63188082	1	10	89
11	0.629511489	0	-1	88
12	0.573017641	1	10	98
13	0.490037283	0	-1	97
14	0.486850139	0	-1	96
15	0.462007181	1	10	106
16	0.409171136	0	-1	105
17	0.371180162	0	-1	104
18	0.291568654	0	-1	103
19	0.265995138	0	-1	102
20	0.214889902	1	10	112
21	0.214028865	0	-1	111
22	0.15727196	0	-1	110
23	0.142752564	0	-1	109
24	0.106523425	0	-1	108

So, this was the data so, the steps that we followed that we sorted the probability estimated probability scores, from you know in the decreasing order starting from the highest values to the lower values. And the actual class was also given in the 1 more column and then we had net value because this was the case where we were incorporating the net value all right, and we were plotting the curve accordingly.

So, net value and cumulative value we used to compute these columns, and then that these values were used to plot the lift curve the cumulative value column and the serial number column. So, this is how we used to do this, now if we look at what we need to change here is the multiple we need to multiply in 1 of the steps in the intermediary step that we need to multiply the net value of a record with proportion of class 1 records in original data.

So, net value of a record for example, if we look at this particular column number a for d column net value. So, there we need to multiply this value by this ratio this proportion of class 1 records in the original data. So, that can now be used to compute the new values, and then those values can further be used to compute the cumulative value, and once we have those cumulative values we can plot our lift curve using on over sample valid validation partition.

Now that lift curve would be adjusted for this over sampling effect, and we would be able to find out the effectiveness of a particular model. Now there are a few more things that we can discuss in it especially in a 2 class scenario right, 1 is that sometimes we might want to have you know you know some records would be there which might not be appropriately or correctly classified by our models. So, can we have some other way

to overcome this problem. So, the records which are difficult to classify by our model we can labeled with a third class option cannot say.

Now, once this kind of modeling is done for all the records. So, most of the records they would be classified as class 1 or some other records would be classified as class 0, the few records which are difficult to classify they can be labeled as cannot say, and then expert judgment can actually be used whether to classify them as once and 0 right. So, this kind of configuration can also be used in a 2 class scenario. So, we will stop here and we will continue our discussion on performance matrix in the next lecture.

Thank you.