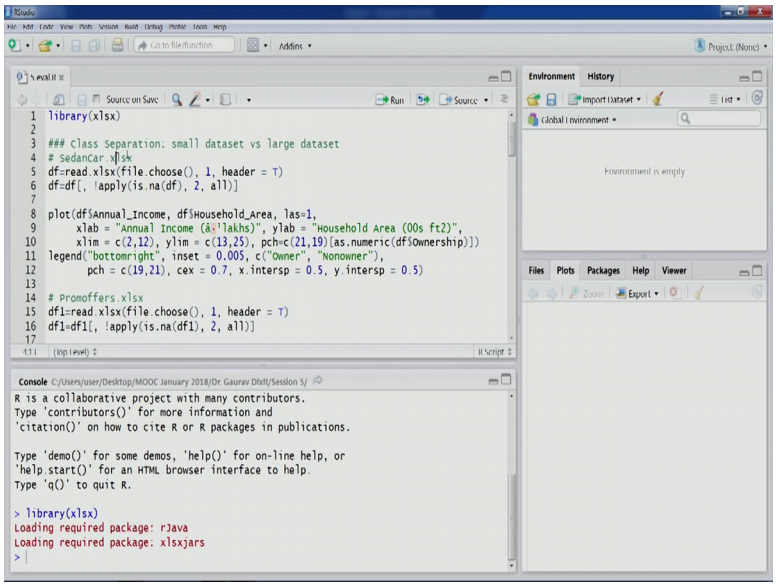
Business Analytics & Data Mining Modeling Using R

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

Lecture – 17
Performance Matrix – Part II ROC/Cut-Off Value

Welcome to the course business analytics and data mining modelling using R. So, last time we started our discussion on performance matrix. So, before we proceed further I will do some exercise exercises that are related to what we discussed in the previous lecture. So, let us open our studio.

(Refer Slide Time: 00:40)

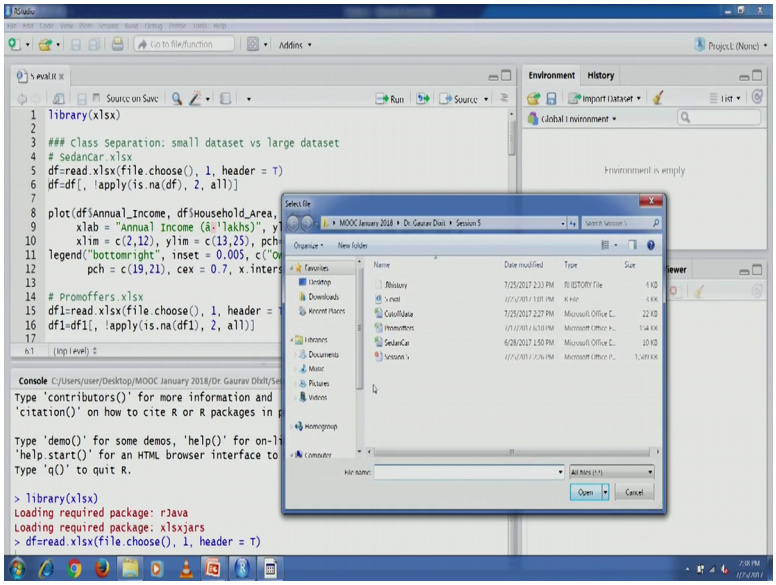


So, as usual let us load this particular library xlsx. So, that we are able to import the data set from excel file.

So, again for this particular exercise we are going to use a sedan car xls data set, data set the particular concept that we want to discuss through this exercise is class separation. Where we, when we started our discussion on class performance sometimes depending on the data set, sometimes for some data sets it might be easier to apply different techniques and get the accurate classifications for different observation, sometimes it might be very difficult because of the data set.

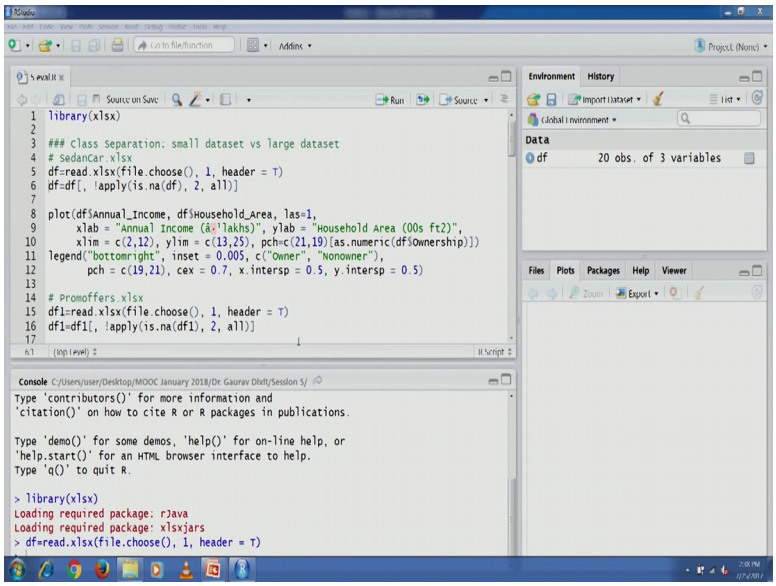
So, let us see this through an example. So, will import this particular data sedan car let us run this code.

(Refer Slide Time: 01:38)



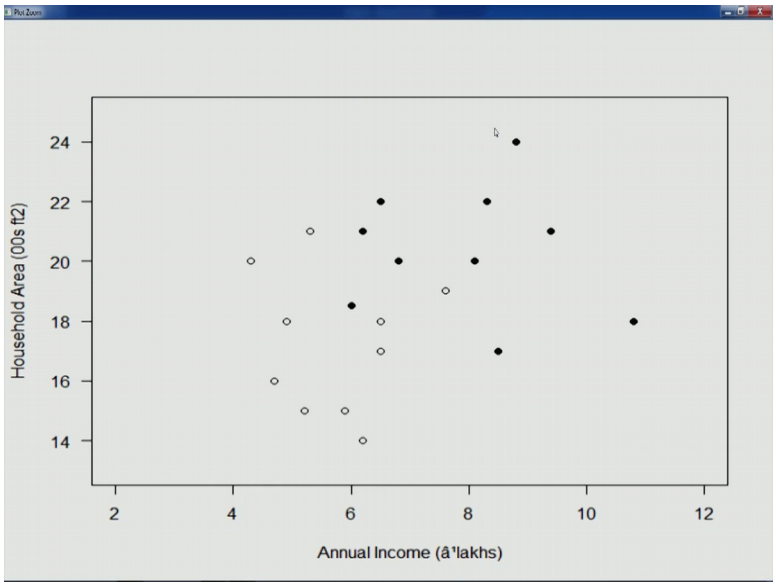
So, this is the data set.

(Refer Slide Time: 01:43)



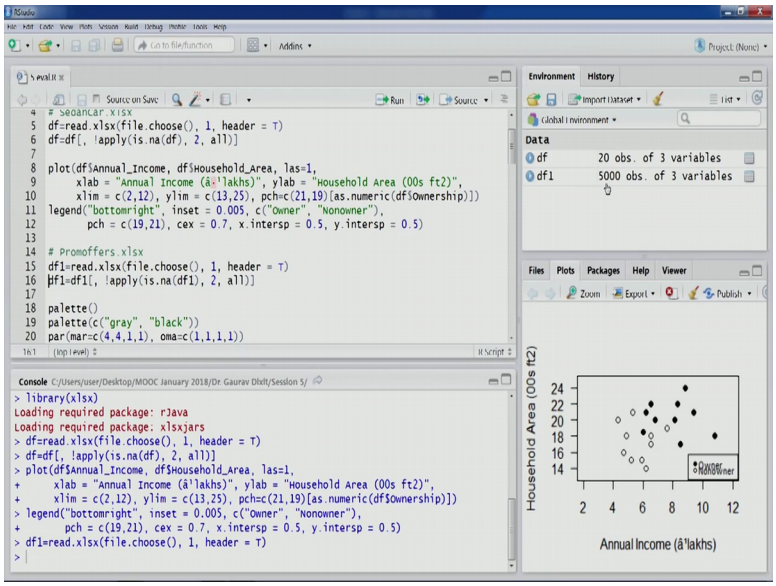
So, you can see in the data set has been imported 20 observation, 3 particular variables let us remove the n a columns and we have as we all, as we are already familiar with this particular data set. So, we are going to plot these 2 values annual income and household area as we have done before as well let us plot this.

(Refer Slide Time: 02:09)



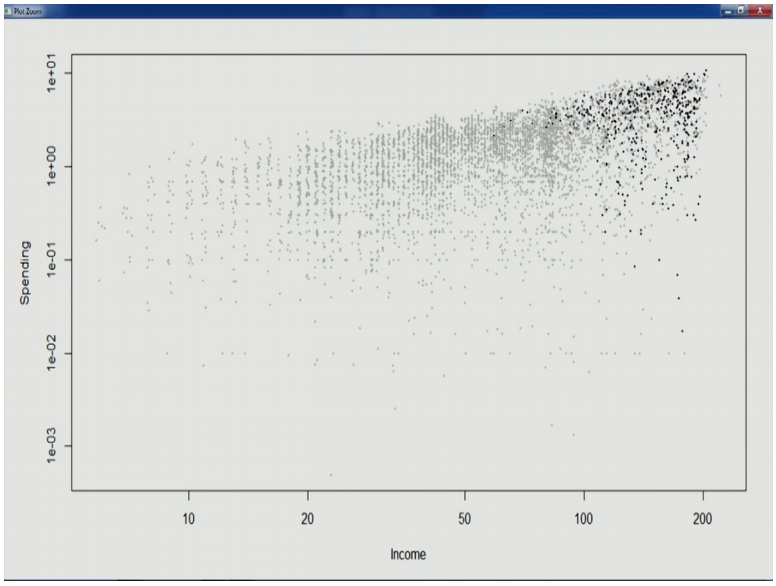
Now, this particular plot we have created in previous lectures as well, now important point to note here is despite being a small data set or maybe that is the that is my times the region also that the classes, there are 2 classes in this data set owner and non owner classes and there is far clear separation between these 2 classes. So, therefore, it is easier for us to apply different methods applied you know build the different candidate models and then a well with the most useful one or to find out the best performing model, right. So, in this particular data set the job is much easier, but sometimes this situation scenario might be difficult, might be different for example, in other data set that we have used before as well the promo offers data set this particular file. So, let us import this one, let us remove the n a columns, let the data just data has been loaded five thousand observation of 3 variables.

(Refer Slide Time: 03:19)



So, earlier data set was 20 observation, 3 variables this one is 5000 observations. So, much larger data set, now in this let us remove the n a columns and we are going to plot this particular data set, this plotting we have done in our previous lectures. So, palette like the last time we had used we had used this gray and black palette that we created, let us change the margin and outer margin settings and for this particular these 2 variables income and spending.

(Refer Slide Time: 04:02)



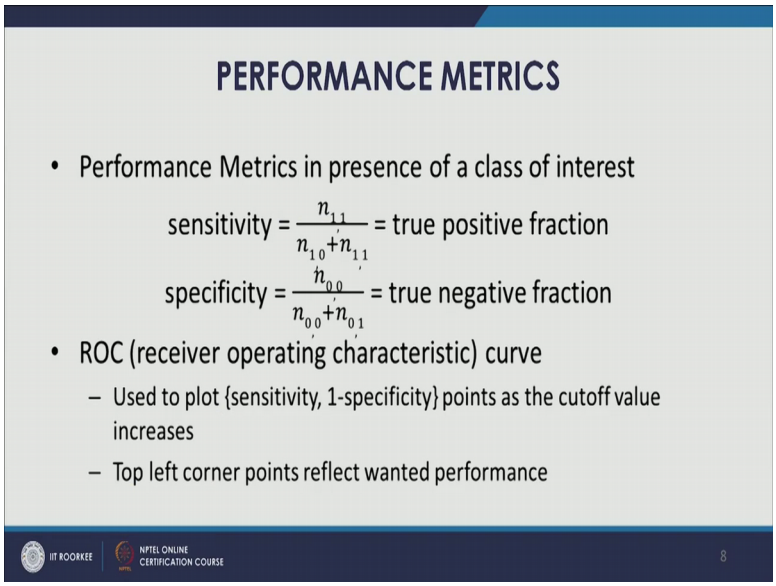
Now, if we zoom in to this particular data set and if we look at other things the points belonging to different classes you would see that this particular region this, left you know this left part of this rectangular region blotting region there is homogeneity there is most of the points belong to or 1 class, but if we look at the right part there is not much clarity, we have points belonging to class 0 and we have point belonging to class 1. So, both classes are present to the separation between classes is not very much clear therefore, when we applied different candidate models on this particular dataset the performance will have to be, will have to be evaluated much much closely and we might not get much improvement in comparison to the benchmark cases.

So, for a classification task point me, for a classification task the class separation is quite important in the previous lecture we also talked about the classification matrix. So, one simple example that I had written here is this creation of this classification matrix. So, this is all a few of our demonstration purpose if you want to create a matrix classification matrix this is how it can be done. Otherwise if you have, if you have information in 2 variables and they are factor variables then table is the command that could be used to generate the classification matrix.

So, in the previous session we talked about error and accuracy 2 2 matrix. So, this is how we can compute them. So, in this particular case you can see the 0 classified as 1 and 1 classified as 0. So, these 2 numbers would actually give the number of miss classification and this would this particular code will compute the error, similarly accuracy club records 0 class 0 classified as class 0 and records of class 1 classified as class 1 and this will give us the accuracy number.

So, let us go back to our discussion on the performance matrix. So, in the previous lecture we stopped at this point and so let us start from here. So, when there is a special class of interest then performance matrix accuracy and error might not be suitable. So, in that case because we have one special class of interest and we might not be interested in other class classification, the misclassification error and whether it is on the slightly higher side or lower side, our focuses on one class in those scenarios we can use these 2 matrix sensitivity and specificity.

(Refer Slide Time: 07:15)



So, in this case sensitivity is about identifying true positive fraction. So, the cases the cases belonging to class 1 how much a particular model is capable enough to identify such cases, similarly specificity measures the capability of model in terms of identifying or removing the observation belonging to class 0, as class 0.

So, that is true negative fraction. So, different candidate models can be compared using these 2 matrix. So, if you look at the name it is names of these matrix sensitivity. So, whether our model is sensitive enough to identify the true positives right, that is a class 1 observation. So, how much of the what proportion of the class 1 class 1 observations are being identified by the models, that would be captured in sensitivity. You can look at the formula as well this is the number n 1 1 divided by n 1 0 plus n 1 1. So, that means, number, number of observation, number of class 1 observation identified as class 1 observation divided by total number of class 1 observation. Similarly, specificity if we look at it is true negative fraction wherein the formula is n 0 0 divided by n 0 0 plus n 0 1; that means, the number of class 0 observation identified as class 0 observation out of total class 0 observation. So, whether the how much, how much capability of the model can be tested can be evaluated using a specificity is in terms of identifying the true negatives whether we are able to eliminate the 2 negatives through our model or not.

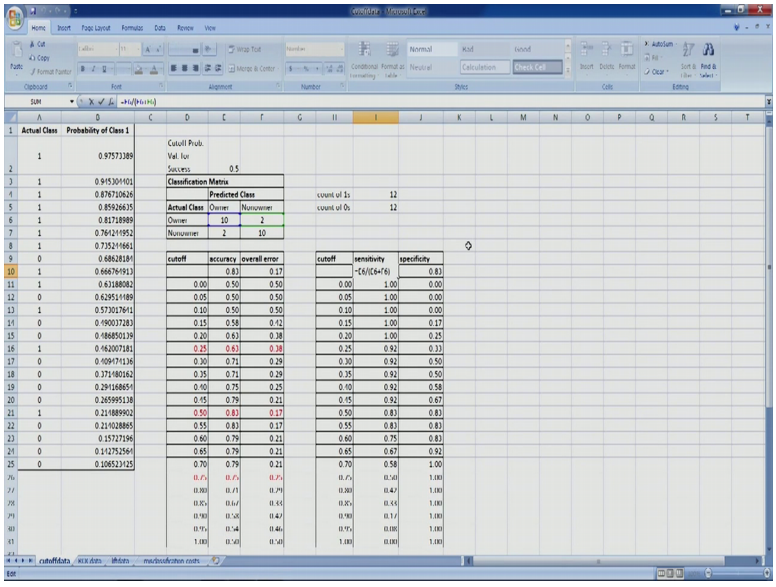
Now, that brings us to our next discussion on ROC curve that is a receiver operating characteristic. So, this product curve is generally used to plot sensitivity and 1 minus

specificity points as the cut off value increases. So, for different cut off values we try to compute these 2 matrix sensitivity and then 1 minus specificity and then we try to plot them for different cut off values.

So, now, the this particular curve is in a way, the way this particular curve was earlier used for radio signals in world war 2 wherein the radio signals and the signals that were received whether a particular signal is identifying a enemy ship or tank that. So, that was identified through a blip in the screen. So, that is where this name is coming from receiver operating characteristic, but this particular curve can be used is being used in multiple domains especially in and to solve analytics problem and especially statistical modelling or data mining modelling is being done.

So, top left corner of this particular plot as we will see through an exercise reflects the required performance, the desired performance that we want from our model. So, let us open our excel file that we used in the last session.

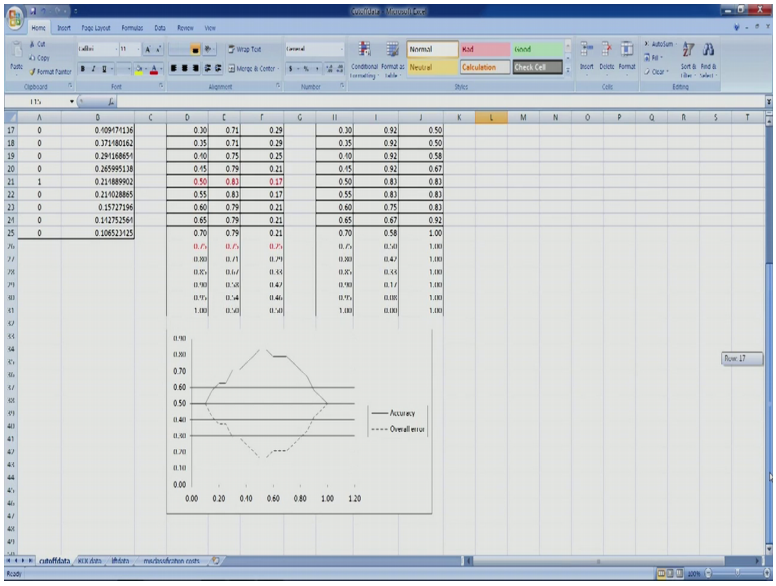
(Refer Slide Time: 11:07)



So, here as you can see in the previous lecture we created this one variable table where we computed accuracy number and overall error for different cut off values, similarly we can compute sensitivity and specificity for different cut off values. So, you can see the values here and you can see how these are being computed.

So, you can see that it is the owner classified as owner that number, number of owners classified as owner divided by a number of owner classified as owner plus number of owner classified as non owner. So, that is the fraction that is how this sensitivity is being calculated, similarly if we look at the specificity this is actually the number of non owners being classified as non owner divided by the total number of non owners so that is how we are computing this. Now, one variable table can actually be, can actually be can actually be computed using these 2 formulas. So, you can see that I have already created this one variable table and for different cut off values you can check these numbers. So, I will be were talking about the ROC curve.

(Refer Slide Time: 12:58)



So, this particular data set can be this particular data that we have just generated can be used to create our ROC curve. So, we have taken out this particular data and copied here in a different worksheet right. Now, this data set will import into R and will create our ROC curve, in the meanwhile the previous table that we had created in the previous lecture this accuracy and overall error.

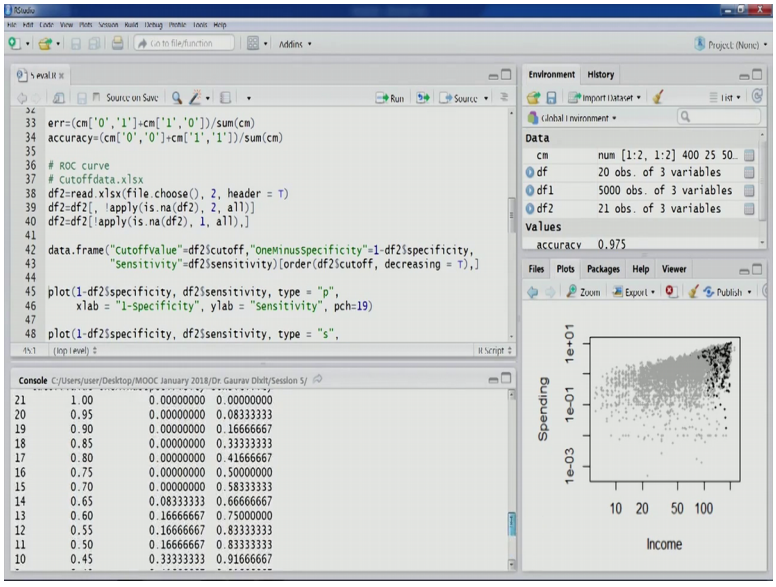
So, one product a plot has been created here, between accuracy for different cut off values if we look at the cut off values, they are plotted in the in the in the x axis and the accuracy numbers and also error numbers they are plotted because error being 1 minus a accuracy they are plotted here in the y axis. You can see for the example that we had discussed in the previous lecture that this is how as we go as we move from where cut off

value of 0 towards around 0.5 and the accuracy keeps on increasing and if we look at this range from 0.4 to 0.8 the accuracy is stable around 0.8 values.

So, a bit fluctuation, but in a way is stable around 0.8 value in this particular region and then again as the cut off value further increases this particular value goes down. So, this is what the data that we had created, the graphic representation of the same. So, now, let us open our studio and will import this sensitivity and specificity data that we have just created. So, let us import this file, you can see 20 one observation 3 variables.

So, let us remove the na columns and also na rows here so that has been removed now what we are interested in we are we want to create an ROC curve. So, ROC curve is with me sensitivity and 1 minus specificity. So, therefore, let us compute that so the data frame that we have is. So, we have cut off value then this number 1 minus specificity and then sensitivity. So, this particular data frame has been ordered by cut off value. So, cut off value.

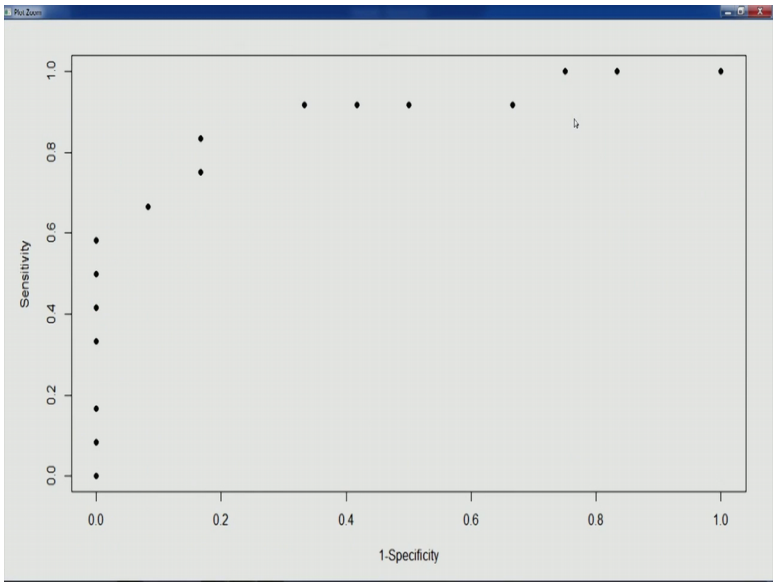
(Refer Slide Time: 15:38)



Let us execute this code, we look at the output you would see that is starting from cut off value of 1 and then it as the value goes down different numbers for sensitivity and 1 minus specificity have been created have been computed.

Now, these numbers would be plotted to create ROC curve. So, let us create this particular plot. So, as we talked about that ROC curve plots, ROC curve plots.

(Refer Slide Time: 16:17)



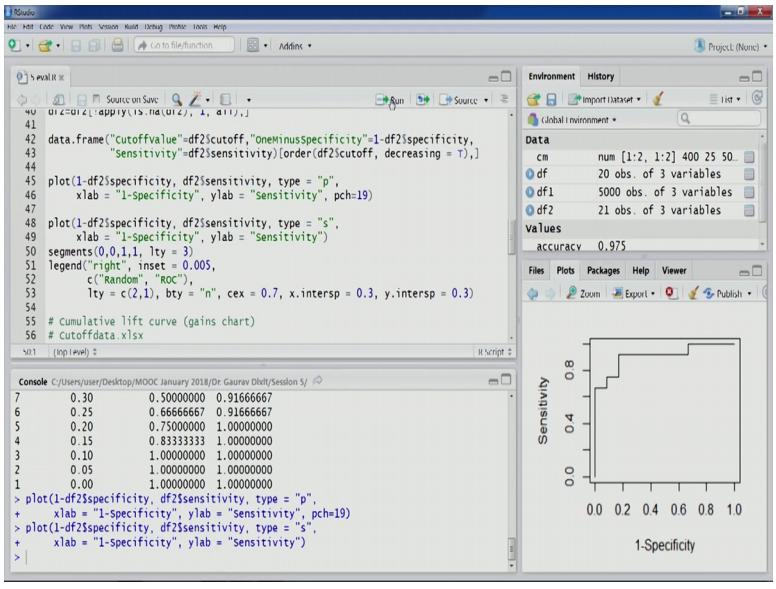
So, the sensitivity and 1 minus is specificity points you can see this has the be look at the again, if we look at the output again for cut off value of 1 and both these values are going to be 0 and 0. So, therefore, you can see the first point is here then for cut off value of 0.95 you can see this value is x value is 0 and the y value is 0.83.

So, therefore, you see this similarly for as we this particular number decreases you would see that sensitivity value keeps on increasing and that is reflected in the plot. So, therefore, as the cut off value moves towards 0.5 from 1 from the value of 1 the model improves in identifying the true positives. So, more true positives are being identified being identified as we move cut off value from 1 to 1 towards 0.5 as this cut off value decreases from 1. So, the sensitivity value has been improving right, as we move further right as we move further you would see that from here if we move further you would see 1 minus a specificity, this value is again some value is there and you would see that this value is also increasing as we move further. So, that means, 2 negatives, 2 negatives as we discussed, 2 negatives as we discussed more false negatives are being identified right in this particular model in this particular plot. So, as we move further the sensitivity improves further. So, as we keep on changing the cut off value sensitivity moves, some sensitivity increase further, but at the expense of false negative values. So, we gain we gain in terms of identifying more ones, but that comes at the expense of misclassifying more 0s so, but the overall sensitivity keeps on increasing.

So, we are interested in this top left corner this is the desirable this is the desirable performance of a model that we want. So, because we are interested in identifying more ones, therefore, top left corner we are interested in this particular region. If we look at the particular 0.5 value at what which point is actually reflecting that value you we can see that, 11 value is reflecting the 0.5 cut off value. Wherein this true negative false negative rate is 0.16 and we can, we can see that sensitivity is 0.83 at 0.5, if we look further than the maximum, we look further then the false negative rate increases quite significantly.

So, therefore, as we talked about we would like to identify the model which is in this top left region and not go more into the right direction. So, therefore, this is the model corresponding to a 0.5 cut off value, if we look at the number of 0.16 is on the x axis 0.16 on x axis. So, it would be I think this point and then you have 0.83, yes. So, this is the point corresponding to 0.5 cut off value and lies in the top left region. So, probably for the cut off value of 0.5 we are getting good a good enough performance even for you know using these matrix sensitivity and specificity to have a. So, these were the points to have the actual ROC curve that is generally used, we can change the type of plot from point plotting to step plot plotting.

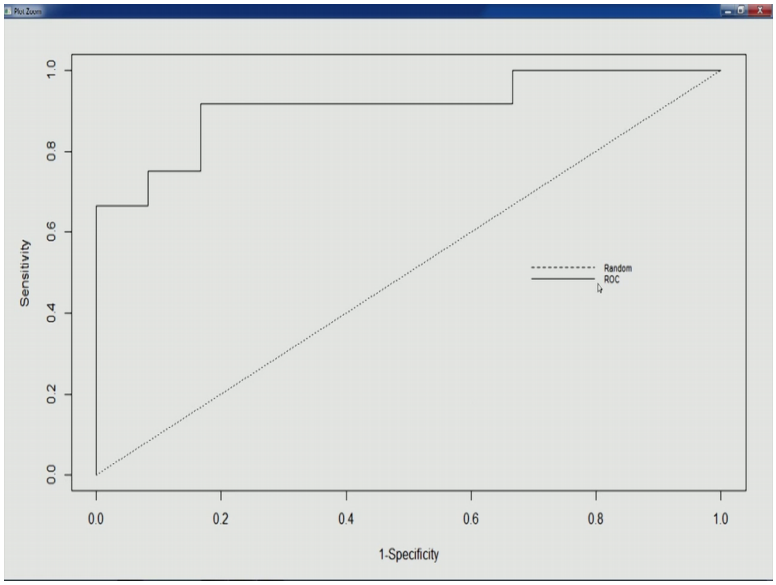
(Refer Slide Time: 21:02)



So, generally step plots are used for ROC curves this is the plot you can see till at till few cut off values few changes in cut off values starting as it decreases from one, the

sensitivity keeps on decreasing sensitivity keeps on increasing as and as the cut off value decreases further, right.

(Refer Slide Time: 21:21)



There is some false negatives are also, some false negatives are also being identified by the model classified by the model right as we move further, further we see some jump in sensitivity, but more false negatives are being classified by the model. So, this is the ROC curve that we want we can draw the reference line. So, this is the reference line representing the average case scenario.

So, this reference line representing the average case scenarios and this particular line representing the ROC curve. So, let us go back to our discussion, now another interesting point that we can understand from this exercise is that while we want to identify when we have a special class of interest we want to identify more of class 1 members because the idea in a business context is generally for example, whether the customer is going to respond to a particular a promotion offer or not. So, in that case we would like to have all those customers which are having higher probability as estimated by the model and mail them our offer. So, therefore, creation of a rank ordering of records with respect to our class of interest becomes more practical.

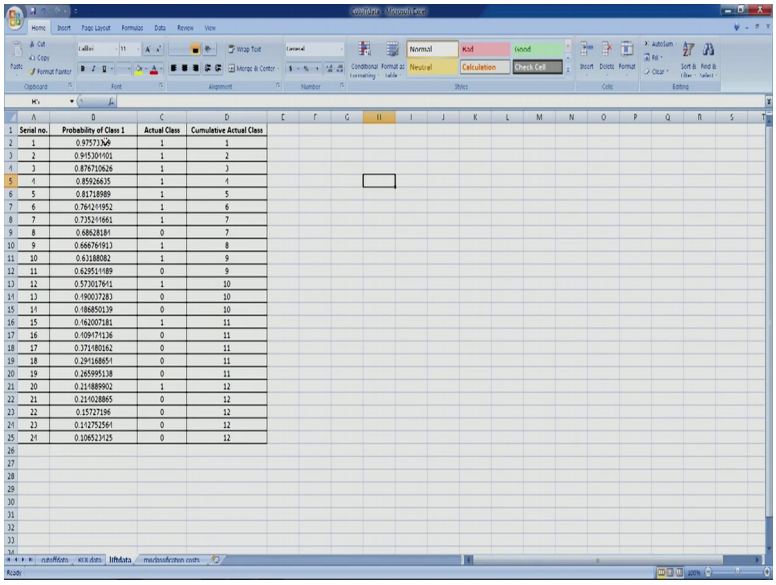
So, how that can be done? How different plots and different mechanisms can be used to do that, to rank on ordering of records for class of interest this could be done based on the estimated probabilities of class membership. So, the records which are having the

highest probability of belonging to a special class of interest they can be taken. So, all those records can we take in first and so that they can be the promotion offering can we send and can be mailed to them first right. So, the lift car, lift curve is there which can actually be used to display the effectiveness of the model in rank ordering the cases right. So, the selected model that we might have we can draw a lift to curve for the same and then find out how effective it is in terms of rank ordering the cases.

So, how it is actually constructed once you have build your model on the training partition you can have your valuation partition scores and using these scores the estimated probability is number we can actually construct our lift curve. Now, the effectiveness of model can generally is, generally is seen using these this particular curve cumulate elliptic curve wherein depending on the probability we look at the cumulative number of records which can actually be, which are actually going to belong to class 1 class of interest this cumulative elliptic curve is also called gains chart. So, this is actually used to plot cumulative number of cases on x axis and cumulative number of true positive cases on y axis.

So, this particular the plot displays the lift value of the model for a given number of cases. So, for a given number of cases the lift value of the model is displayed with respect to random selection. So, if we just rely on the probability value of class membership. So, let us say if there are 20 observation and ten belong to one particular class class of 1. So, therefore, probability of a particular record belonging to the class 1 is going to be ten divided by 20 all right so that is going to be 0.5. So, if we do random selection how much more our model is going to help us in terms of identifying those cases as belonging to class one how much lift in comparison to this random selection this average scenario is going to be given by our model. So, that we can do through cumulative lift chart or gains chart. So, we will do again discuss this through an exercise.

(Refer Slide Time: 26:06)



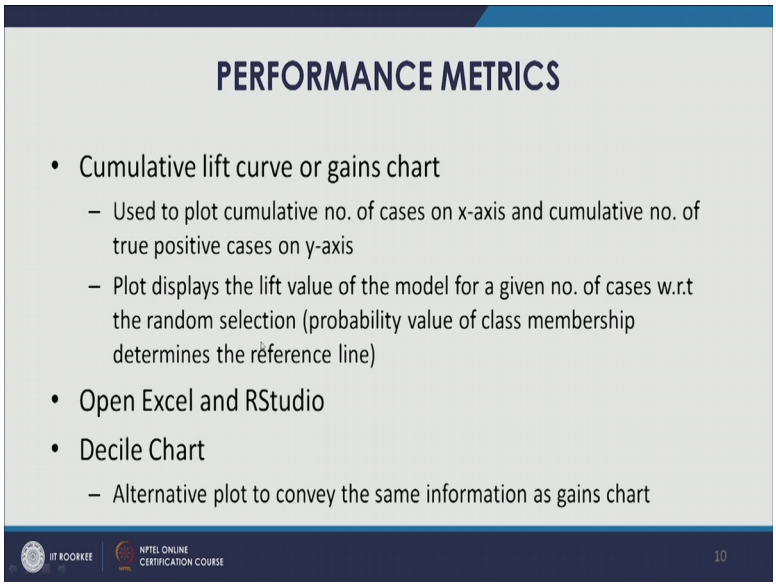
Serial no.	Probability of Class 1	Actual Class	Cumulative Actual Class
1	0.975732	1	1
2	0.915301401	1	2
3	0.876710626	1	3
4	0.85926635	1	4
5	0.81718989	1	5
6	0.764219352	1	6
7	0.735211661	1	7
8	0.68628184	0	7
9	0.666769913	1	8
10	0.6318082	1	9
11	0.629511489	0	9
12	0.573017641	1	10
13	0.490037283	0	10
14	0.486850139	0	10
15	0.462007181	1	11
16	0.409171436	0	11
17	0.371480162	0	11
18	0.294168654	0	11
19	0.265995138	0	11
20	0.214889902	1	12
21	0.211028865	0	12
22	0.15727196	0	12
23	0.112752564	0	12
24	0.106523425	0	12

So, let us open our excel file, now if we look at the this particular same data, the same 24 observation and their estimated probability values can be seen here, we also have actual class corresponding to each record right. So, what we want to compute is the cumulative actual class. So, for now let us look at this. So, the these probabilities are pre arranged from. So, in decreasing order. So, the record with the highest probability or belonging to class one is first and then it is followed by records with slightly you know in decreasing order having probability values in decreasing order and then the actual class membership is there.

So, if we look at the cumulative actual class for the first record we have 1, once we get the second record which is also then the number the cumulative number of records belonging to actual class will be 2. So, in this fashion will continue till we have actual class as identified as 1 now when we come to the first record which is misclassified 0, but if you look at we look at the probability a value it is more than 0.5 which actually 0.686, but it has been miss, but it is going to be misclassified, but it is going to be classified as a record long to class 1, but it is actually class 0. So, we do not add this particular. So, this there is not going to be no addition in the cumulative actual class this will remain 7 then in the same fashion will give on and will keep on accumulating these numbers.

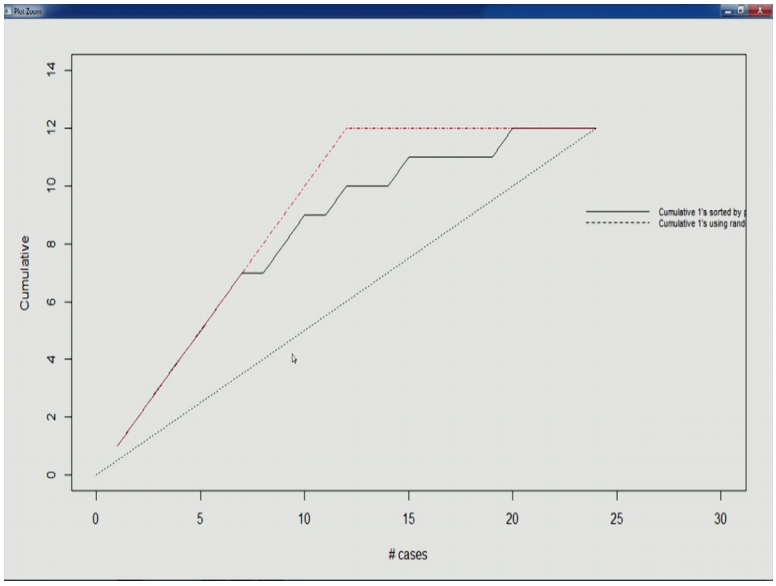
So, once we have prepared this particular data then we can go ahead and create our plot.

(Refer Slide Time: 28:12)



So, let us open our studio. So, first let us import this particular data, you can see 24 observation, 4 variables. Let us remove the na columns let us look at the range of this particular new variable cumulative actual class and range. So, we have we know that we have 24 observation. So, the plot is going to be between number of cases on x axis and the cumulative actual class number on y axis. So, let execute this code and you would see the plot has been generated.

(Refer Slide Time: 29:09)



Now, let us also plot the reference line. So, that being that is the reference line. So, now, from this, from this ref reference from now this particular lift this particular cumulative lift curve or gains chart has been generated. Let us also create the legend and few other lines, now let us look at the plot now you can see this is the dotted line represents the cumulative ones using random selection. So, if we randomly select and rely on probability value this is the line then the actual line is cumulative one sorted by predicted values right. So, we will stop here and will continue our discussion and try to interpret this particular gain chart in the next lecture.

Thank you.