Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

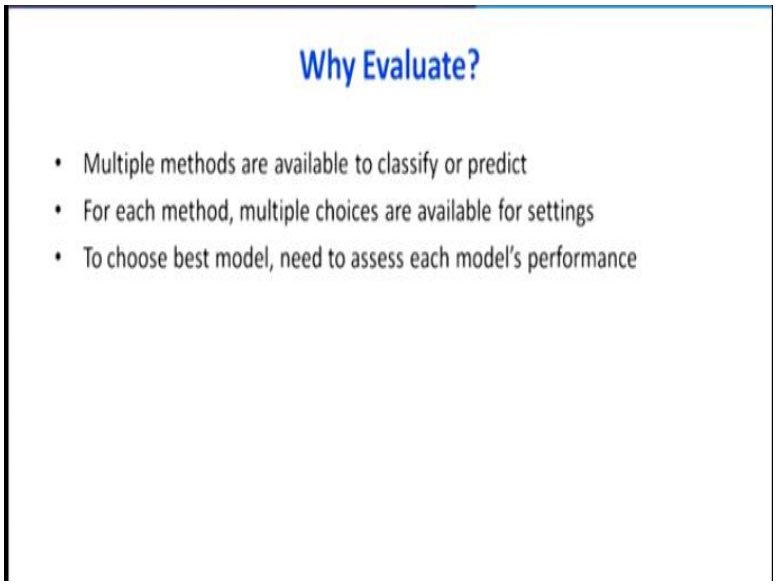
Lecture – 41
Confusion Matrix and ROC

In this class, we are going to talk about how to check the performance of a logistic regression model. There are 2 ways to do that one; one is checking confusion matrix, another one is ROC, we will explain what is this confusion matrix and ROC, then we are going to see how we can check using these 2 criteria that the model which we developed is good or not.

(Refer Slide Time: 00:54)

The agenda for this class is we will see what is is confusion matrix and receiver operating characteristics curve.

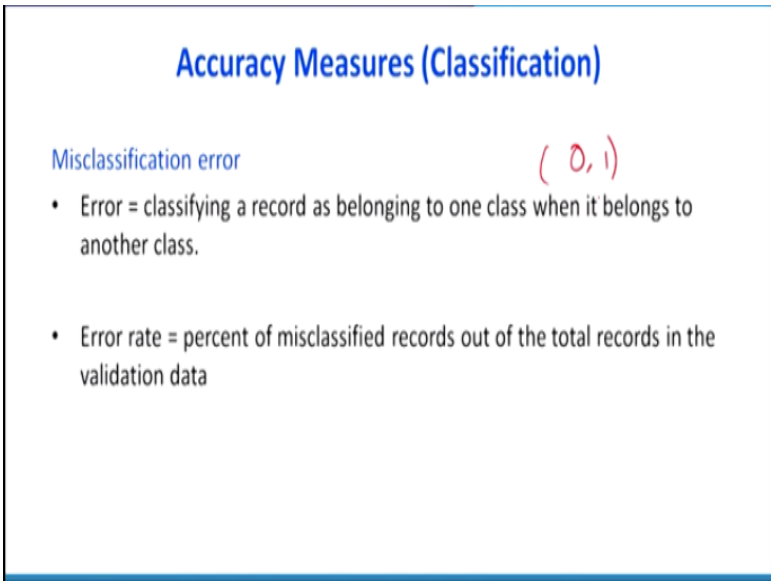
(Refer Slide Time: 01:01)



Because we have seen in our previous example, there may be a different method to classify a set of data set. One of the methods is our logistic regression that is used to classify to into 2 category, whether it is 0 category or 1 category but we want to see which method is the best one, so multiple methods are available to classify or predict. For each method, multiple choices are available for setting.

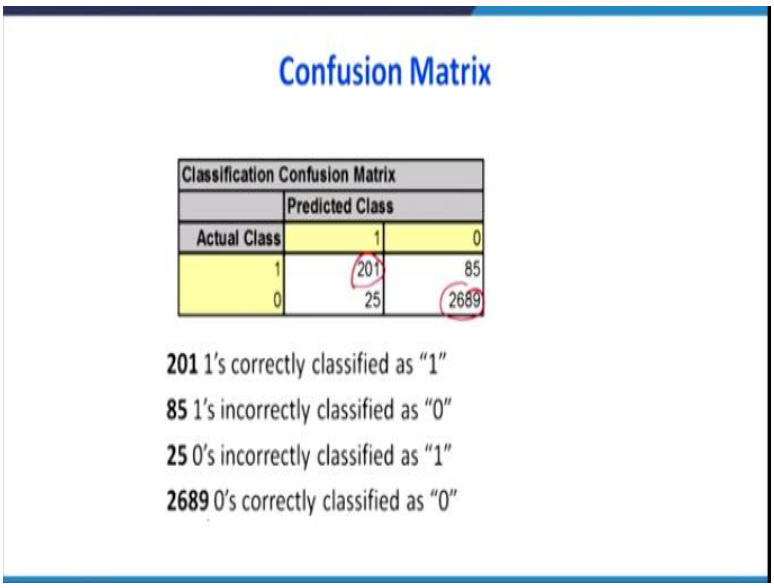
Here, multiple choices means that threshold value which we are going to say that beyond this probability, you should go to 1, below this probability you should come to 0, so that is our multiple choices. So, we have to know to choose the best model, we have to assess each model's performance that we will see in this class.

(Refer Slide Time: 01:53)



In the classification context, how to measure the accuracy; one term is misclassification error, first we will see what is error. Error is classifying a record as labelling into one class when it belongs to another class. Suppose, when we say 0 to 1; 0, 1, there are 2 category; we can predicted, sometime what will happen, we may wrongly predicted, instead of saying 1, we will say it as 0, instead of saying 0, we will say it 1, so that is error. Error rate is percentage of misclassified records out of the total records in the validation data that is an error rate.

(Refer Slide Time: 02:42)

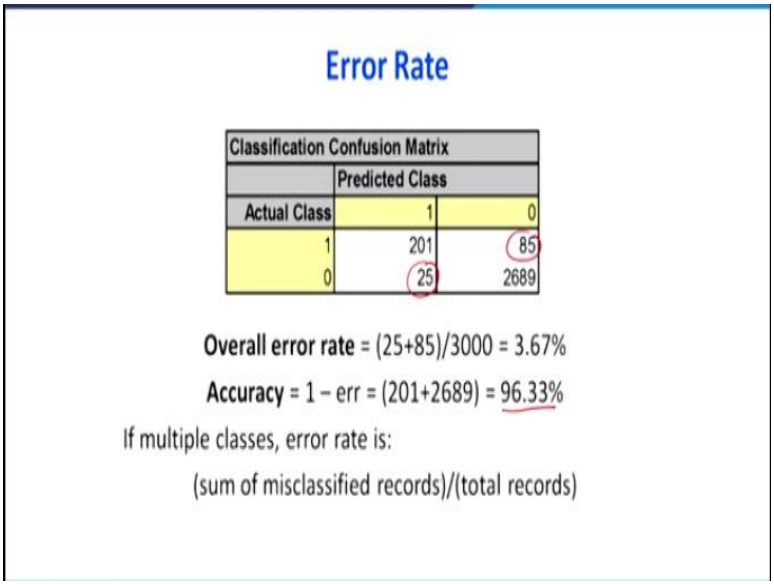


This is an example of confusion matrix; you see in row there is a actual class, in column, there is a predicted class. So, in row we see 1 0, 1 0, if you can predict 1 1 that is a correct one, actual also 1, the predicted value also 1, so like that we got this many number of data set. The other

possibility; the actual is 0, the predicted also 0, so these 2 columns, 2 cells are the correct value. So, here the frequency of correct saying 1, when it is actually 1 is 201.

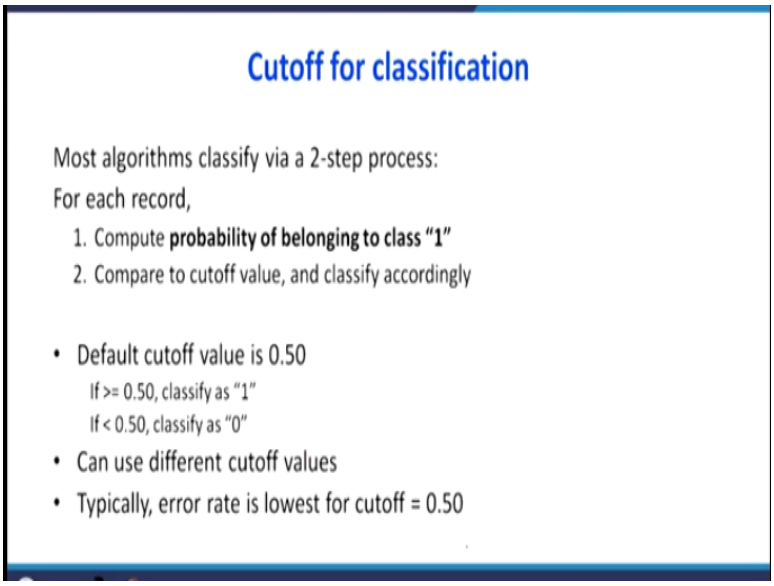
The frequency of saying 0, when actually 0 is 2689, so the 201; 1 is correctly classified as 1, here the 85, 1 is incorrectly classified as 0, actual is 1 but we are predicting 0 that is your 85, the 25 represents incorrectly classified as 1, actually it is 0 but we classified as 1, 2689's are classified as 0, actual also 0, the predicted value is 0, so this is the set up for confusion matrix. This matrix is useful to find out the accuracy of our regression model.

(Refer Slide Time: 04:07)



We go here, how to find out the error rate; from error rate, we will see how to find the accuracy of our predicted model. See the overall error rate is; there are 2 error possibility, this 25 and 85, when you add this 25 + 85, the overall data set; overall count is 3000, so the error rate is possible error divided by 3000, so 3.67, the accuracy is 1 – error rate, so 1 minus that will give the 96.33%. If multiple classes is there, here only there are 2 classes there, one is 1 0. Sometimes there will be possibility of 2 also; in that case the error rate is sum of misclassified records divided by total records.

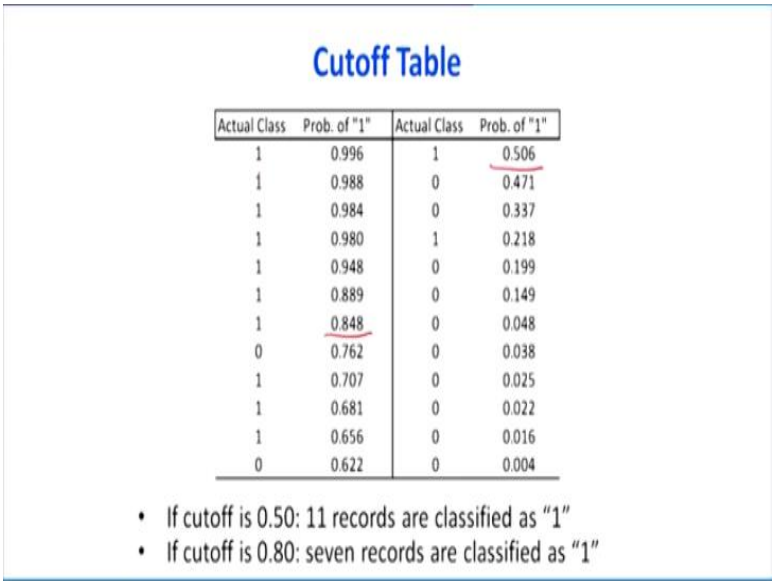
(Refer Slide Time: 05:02)



Here, we will see cut-off for classification, so we need to have cut off to say when it is 1, when it is 0, most algorithms classifying via 2 step process. For each record, compute the probability of belonging to class 1, compare the cut off value and classify accordingly. The default cut-off value is 0.5, if the cut-off value is greater than or equal to 0.5, we will classify as 1. If the cut-off value is less than 0.5, we can classify as 0, okay.

In the probability range, we can have below 0.5, we say it is 0, above 0.5 is 1, this is the default value, we can use different cut off values, typically error rate is lowest for cut off, when you take the cut off value is 0.5.

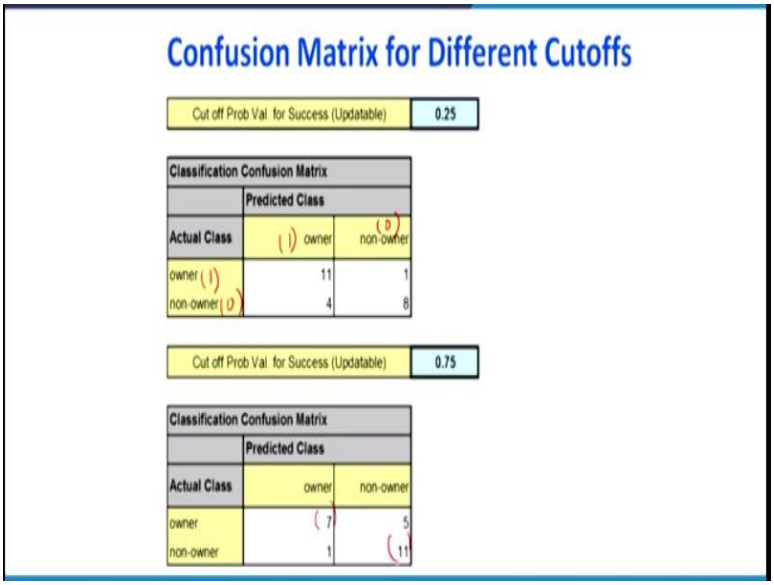
(Refer Slide Time: 05:58)



For example, look at this picture, this is one example of our say, logistic regression model, this is our estimated y value. As I told you in our previous classes, the estimated y value is a probability, 0.996, 0.988 up to this is the continuing of this one. Suppose, if you keep cut-off is 0.5, what is the cut-off 0.5; 0.5 and above we are going to call it as 1, so this category. When you keep the cut-off here, there is 11 records are classified as 1; 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11.

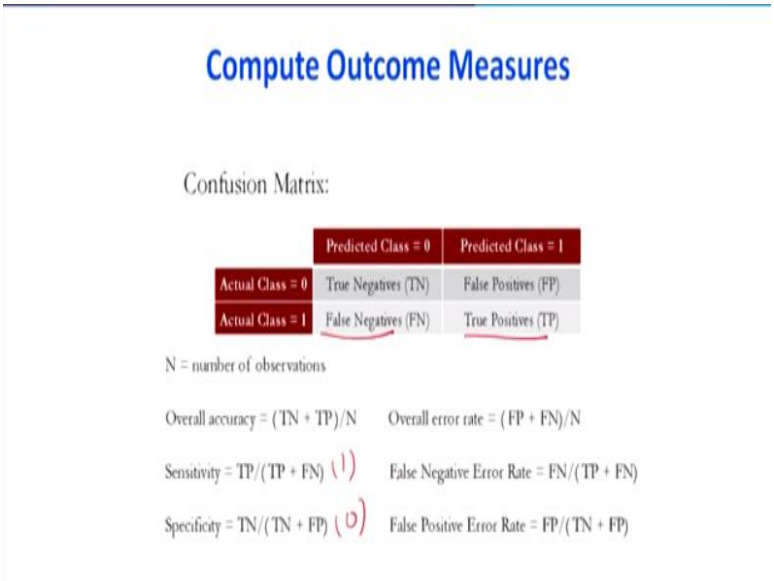
When you cut-off is 0.8, suppose when you put a cut-off here, 7 records are classified as 1, where 1, 2, 3, 4, 5, 6, 7, the problem comes what should be the right cut-off to of classify it is 1 or 0 that is in our hand. Sometime, if you keep very high cut-off that also not good, if you keep very low cut-off that also not good that we will see, what is the meaning of keeping higher cut-off, what is the meaning of having lower cut-off?

(Refer Slide Time: 07:19)



Assume that my cut-off value is 0.25 in our previous problem, so cut-off this can be updated, mini software packages can be we can keep different cut-off, so when you keep cut-off is 0.25, say this value, the actual is owner, the predicted may be 1 also, you can call it as 1, this also 1, this is 0, this is 0. So, 1 1 is the correct prediction, 0 0 it is correct prediction when you are keeping it is a 0.25. Suppose, you increase the cut-off value to 0.75, what will happen; were able to predict only 7, here able to predict 11. So, what is happening; when you update the cut off, we are getting different confusion matrix that confusion matrix, every confusion matrix we will say about the overall accuracy.

(Refer Slide Time: 08:24)



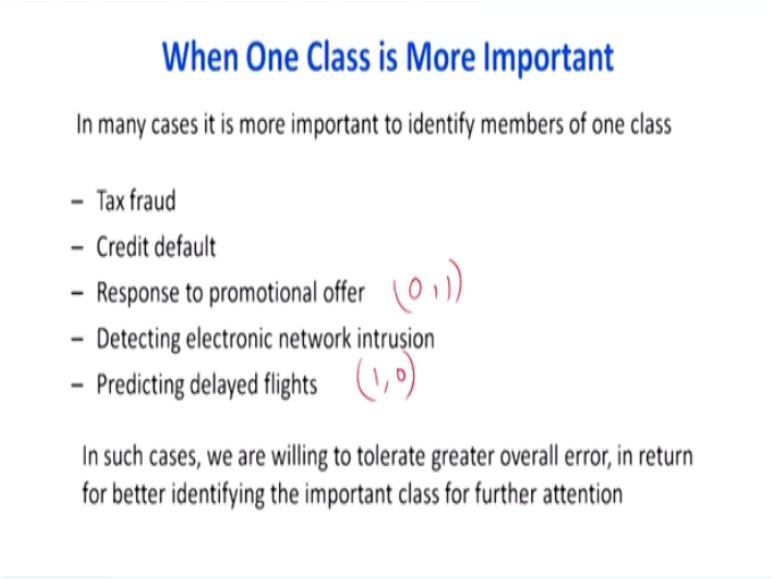
From the confusion matrix, generally the custom is first write 0 then 1, here also 0 1, you see here actual also 0, the predicted value also 0, it is a true negative, you see this diagonal value, actual also 1, predicted value also 1, so it is true positive. Whenever we do this mistake, what is happening here the actual is 0 but we are shown it is 1, so it is a false positive. I have some example in coming slides, what is the meaning of false positive, intuitively you can understand.

Similarly, the actual is 1 but you are predicting 0 that is your false negative, so these 4 cells are used to find out there are different parameter to check the prediction power of our regression model. The overall accuracy; these 2 cells are correct values TN, true negative plus true positive divided by sum of all cells value that is overall accuracy. The second point is sensitivity; sensitivity is true positive divide by true positive plus false negative that is sensitivity.

Because why we call it a sensitivity; actual also 1, predicted value also 1, so that is a sensitivity, so here the context of sensitivity; sensitivity of a testing machine that I will show you in the next slide. Then, specificity, specificity is true negative divided by true negative and false positive that is specificity. Here if you are predicting 0, we call it as specificity, if you are predicting 1 in a right way we are calling it a sensitivity.

Then the next term is overall error rate, what is an overall error rate; the false positive is one error plus false negative is another error divided by total number of elements. False negative error rate, where this is a false negative divided by true positive plus false negative that is a false negative error rate. False positive error rate, false positive divided by true negative plus false positive. I will explain what is the meaning of false positive, false negative in coming slides.

(Refer Slide Time: 10:59)



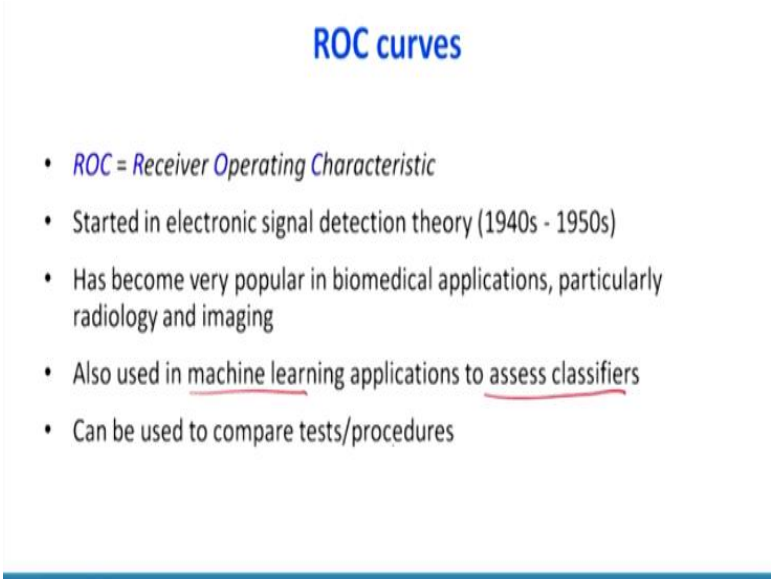
Many times, the accuracy of the model is not important, sometime we may say that the one class is more important for example, predicting 1, when it is actually 1 that is more important. In many cases, it is more important to identify members of one class whether it is 0 or 1 but many time it is 1, 1 means when actual is predicted, actually it is 1, the predicted value also 1, so that is our more important class, we are not bother about when the 0 is predicted as 0, that is not important.

If it is 1, we should predicted as 1, so that time, the only one level is more important, for example tax fraud, credit default, response to promotional offer, detecting electrical network intrusion, predicting delayed flights, so there is a 2 possibility there, a person has done tax fraud or not, credit fault; default there are 2 possibility, this fellow will default or not. Response to promotional offer; whether this person will take the promotion offer or not.

If it is not taking no problem but we are considered about whether he is going to take the promotion offer or not because only between 0 and 1, we are more focus on 1, 0 is not important

for us, detecting electronic network intrusion, predicting delayed flight, whether the flight will be delayed or on time, so we are sometime we concerned about only the on time. In such cases, we are willing to tolerate greater overall error, in return for better identifying the important class for further attention. So, when you want to focus only one class out of these 2, that time accuracy is not important, something else important that will say.

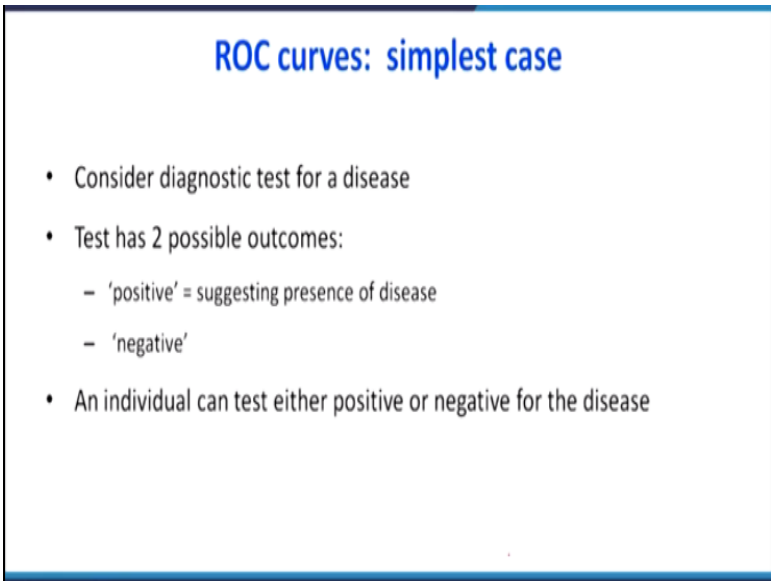
(Refer Slide Time: 12:52)



That is done with the help of this curve called ROC curve, ROC is receiver operating characteristic curve, this curve is used to identify what should be the our threshold value to decide whether this category belongs to 1, whether this category belongs to 0, it was the idea started in electronic signal detection theory in 1940s to 1950s. It has become very popular in biomedical applications particularly, radiology and imaging.

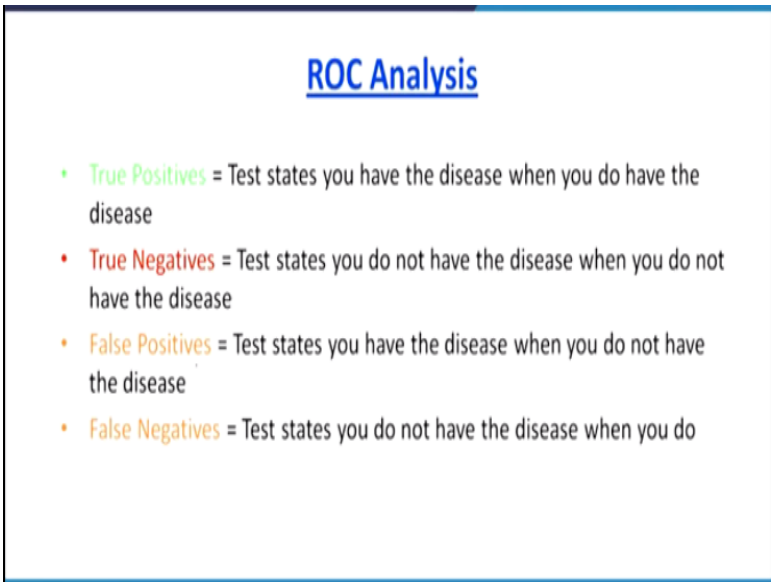
Because if you want to predict a person having a disease or not, so this ROC is more suitable to decide whether there is a difference of different test, also used in machine learning application to assess classifier, in this class this ROC curve is used to decide or to evaluate whether the classifier is correctly classifying or not. Even it can be used to compare test or procedures here in the context of medical. So, what kind of operation can be done so, what kind of operation is more suitable for the patients?

(Refer Slide Time: 14:10)



We will see one example simple case; consider diagnostic test for a disease, you are asked to go for test, say medical test. The test has 2 possible outcomes; one is you may get positive that suggest that presence of disease, you may get negative that says absence of the disease. An individual can test either positive or negative for the disease. There are 2 possibility is there, a person may have the disease but you may get the negative report. Sometime a person may not have the disease but you may get positive report, so that is why the error started to come.

(Refer Slide Time: 14:59)



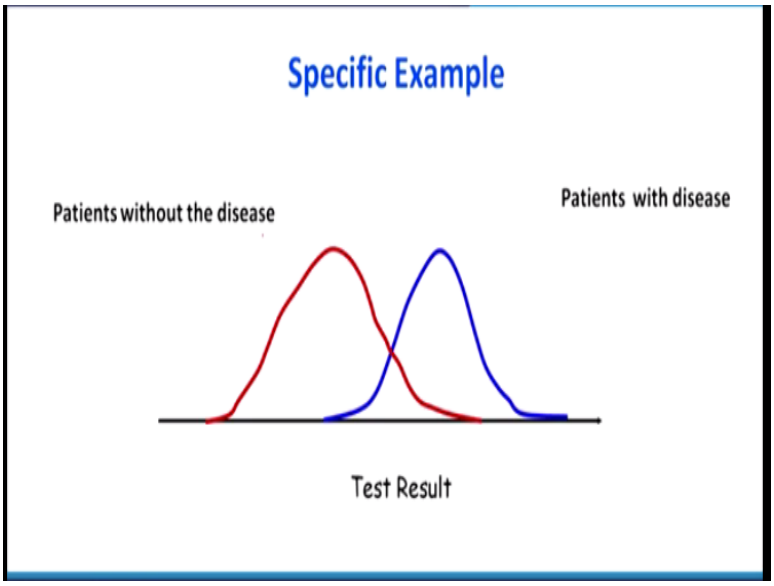
These terms you are going to use so often in coming slides, one is what is a true positive? Pictorially, I will show you in the next slide, the test states that you have the disease when you do have the disease, that means the person having the disease correctly it is saying that yes, you

have the disease. The true negative means the test states that you do not have the disease, when you do not have the disease, this is also no problem, you do not have disease, the report also; the test also says that you do not have disease.

The problem comes here in the false positive; the test states that you have the disease but when you do not have the disease, what is the meaning is that actually you do not have the disease but you shows the positive, positive means that you are saying that there is a disease (()) (15:55) is very dangerous that means, you do not have disease but the test that machine says is no, you have the disease.

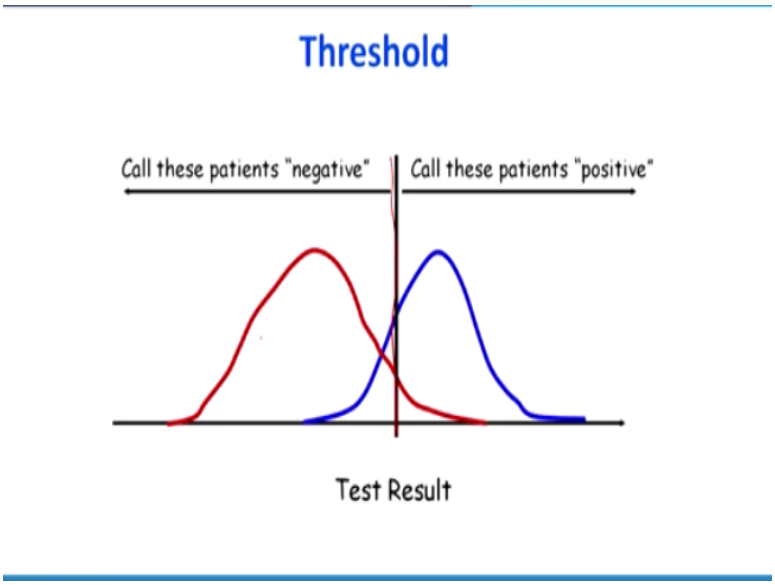
Then the doctor started to, start the medication that may be dangerous also, there may be another category, test states that you do not have the disease, when you do, this also very dangerous actually, you have the disease but the test says you do not have the disease that this fellow may not get the proper medications because the test told that you do not have the disease, so both false positive and false negatives are dangerous.

(Refer Slide Time: 16:34)



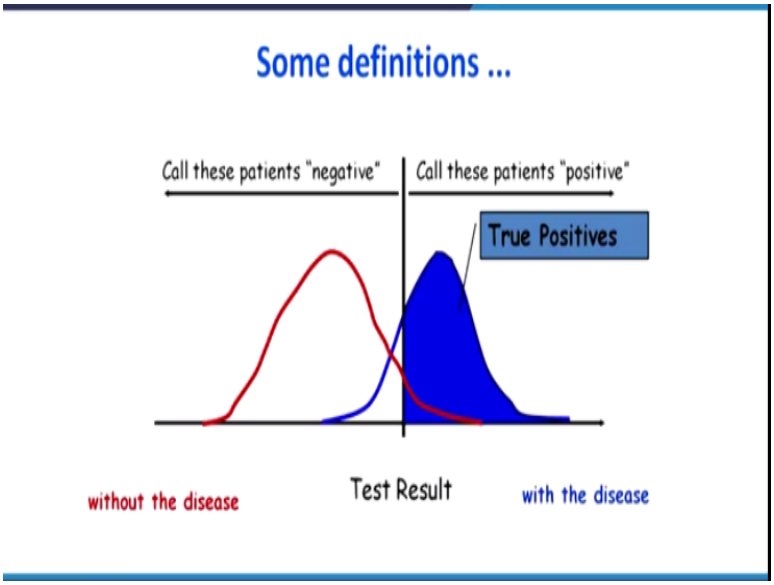
Look at this picture; the red colour shows that patients without the disease, the blue colour shows that patients with disease. Now, this is the test result, there are 2 possibilities there, a person without the disease, with disease.

(Refer Slide Time: 16:55)



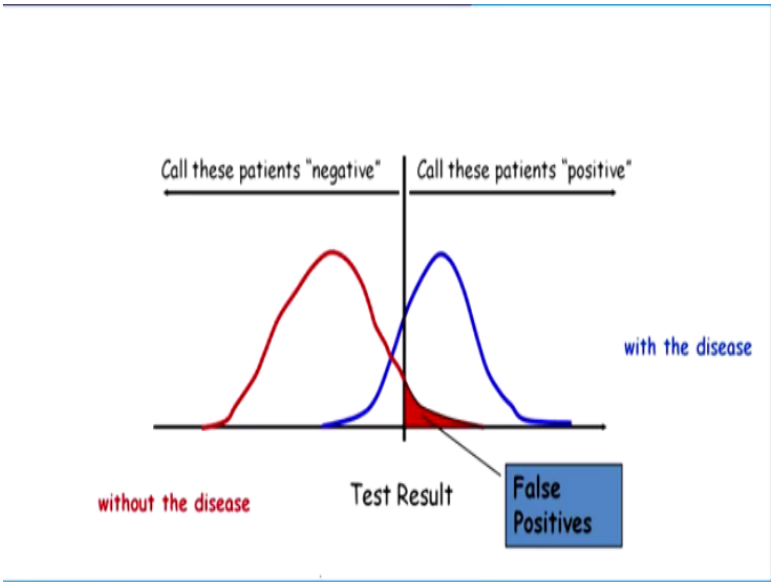
Now, if you keep a cut-off here like this, you see that this in the x-axis shows kind of a probability. So, beyond this right hand side, you can call the patients having disease say, positive, beyond the left hand side of this line, you are going to say that the test shows negative that means that the patient is not having any disease.

(Refer Slide Time: 17:21)



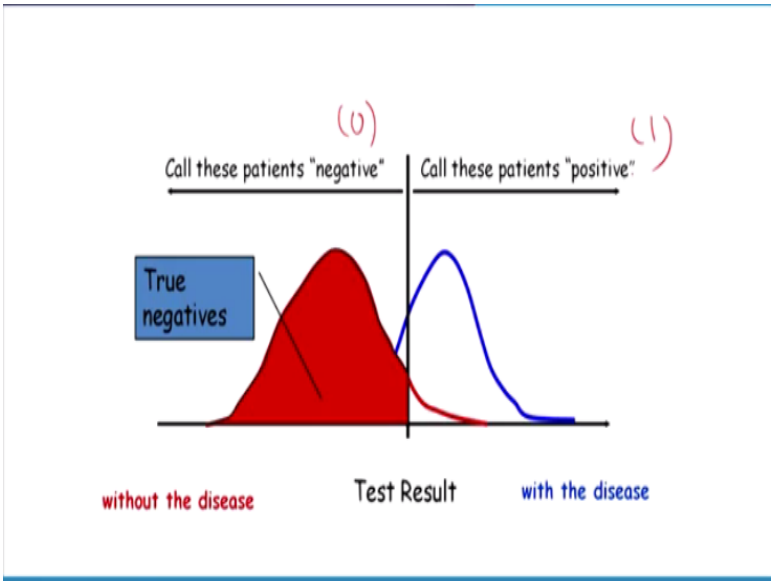
Now, you see that the blue one, this portion says true positive, what is a true positive? The person also actually have the disease, the test also says, yes you have the disease. Suppose, say for example 1 1, so that is a true positive.

(Refer Slide Time: 17:39)



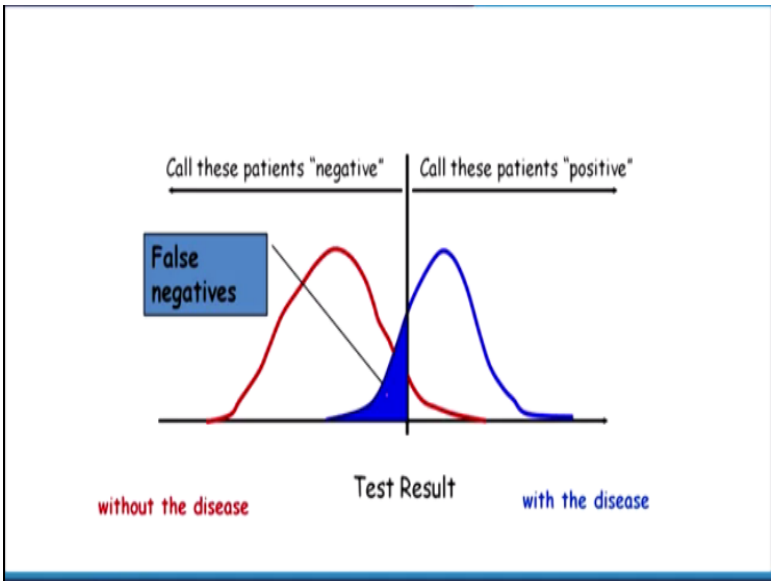
Now, look at this because the both negative and positive there is an overlapping, see the red portions actually, this red portions actually belongs to negative just because it is lying on positive side of this curve, we are going to say it is a false positive. False positive means actually, he is not having disease because of this cut-off which we have chosen it is lying on the positive side, we are going to say false positive, this is not good.

(Refer Slide Time: 18:26)



Because a person is not having disease but you are going to say is a disease, then you see another category true negative. When there is a cut-off, the left hand side portion says that true negative means the person not having disease, the test also says not having disease, it is like 0 0, 0, you code it to 0, no disease, this is disease, this also no problem because we will not bother about.

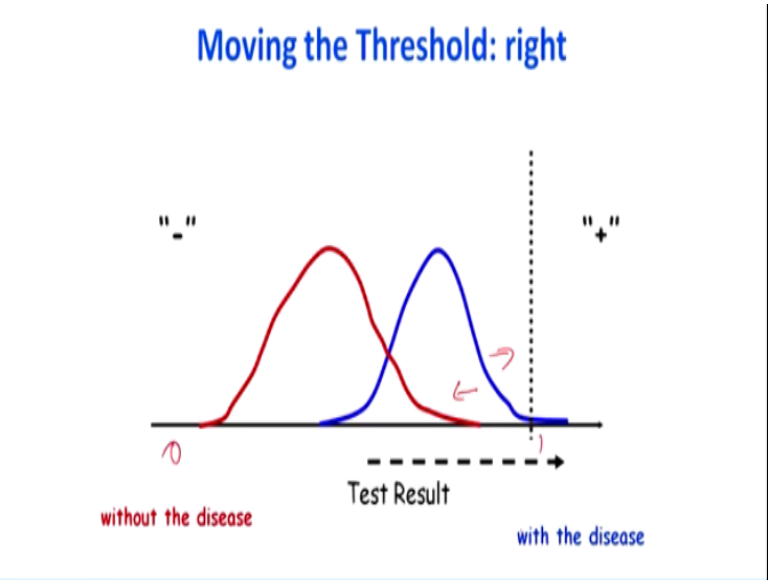
(Refer Slide Time: 18:50)



Now, what will happen in this case, this case actually is a false negative, actually this much portions of the blue they have the disease but it is lying on the negative side of the curve, we are going to say it is a false negative. The very common example for this one is sometime people may have confusion that person is having heart attack or the gastric trouble, so what will happen sometimes this is the false negative.

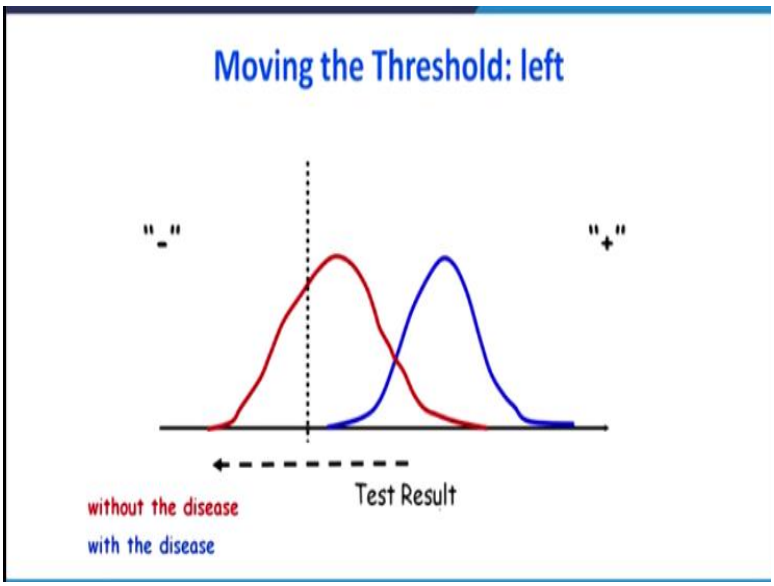
Actually, he had the heart attack, some people may suggest no, no, it is due to gas, so this is a false negative, this also very dangerous. Now, the question comes what should be the cut-off, suppose if you increase this cut-off what will happen?

(Refer Slide Time: 19:36)



That was the next one, suppose we have increase this cut-off like this, so what will happen when you increase the cut-off, so whoever comes there, we will give the report that negative, right because of our, so below this point is negative, below this point is positive. So, whoever goes to the pathological department, they will get a report that you do not have any disease because we have kept higher cut-off. Now, let us see what will happen now, see this one, this is the suppose, since it is the probability 0 to 1, see that whoever goes there, they will get a report, negative report.

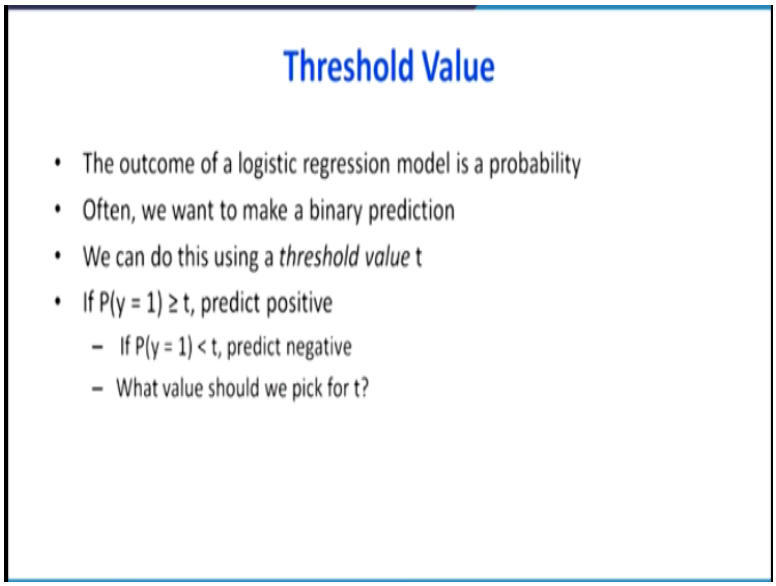
(Refer Slide Time: 20:21)



Then we will see another category, suppose if we decrease the cut-off, what will happen, when you decrease the cut-off, whoever goes to the pathology laboratory, they will get a positive

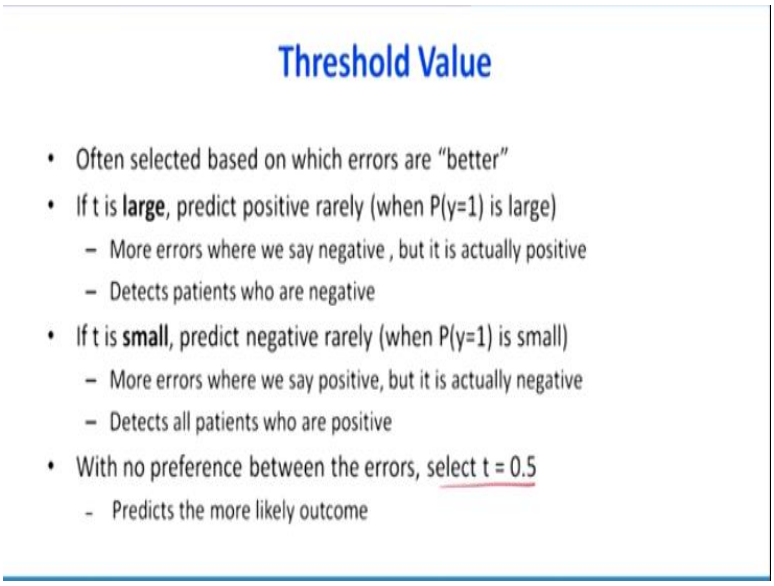
report; positive report means you may not have the disease but the report is going to say that no, you have the disease, you have to start the treatment. So, what is happening, the cut-off value plays very vital role to decide to minimise these errors. So, in this class what we are going to do, we are going to say that what is a role of this cut-off value and the accuracy of our predicted model or the classification model.

(Refer Slide Time: 21:02)



So, we have to have the right threshold value; threshold value means that the vertical line, where it has to be chosen, whether it has to be chosen right hand side or left hand side because see the outcome of a logistic regression model is a probability often we want to make a binary prediction whether it is 0 or 1. We can do this using a threshold value t, call it as t, above this threshold value we are going to predict it is a positive, below the threshold value we are going to say it is predictive. Now, what is happening, what value should be pick for t, what should be the value of the cut-off value.

(Refer Slide Time: 21:41)



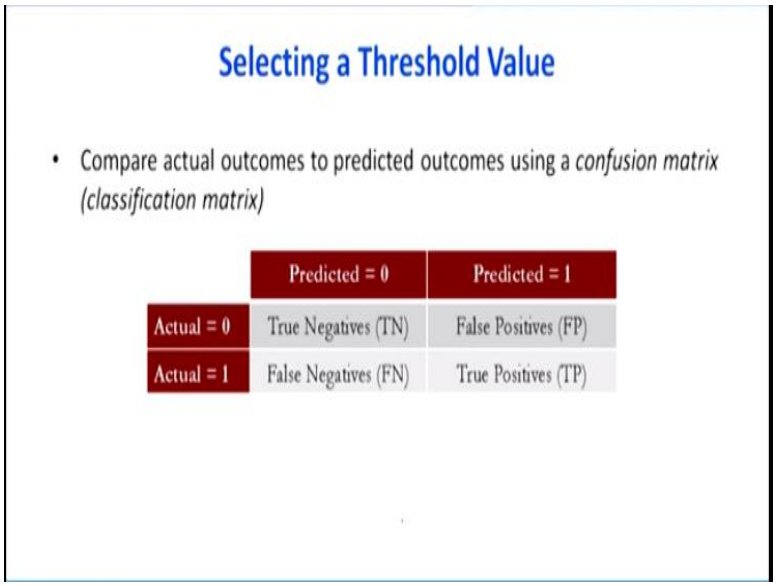
This cut-off value is chosen based on which error is better, there are 2 error we have seen; false positive, false negative. If t is large, which I was shown you previously, predict positive rarely that means, if the t is high, we say report always that it is negative, so more errors where we say negative but it is actually positive, so what will happen here this curve, when you keep the higher threshold value, you see this fellow is a positive, this fellow is really having the disease.

But since because you have chosen higher cut-off value, we are going to give a report saying that negative that also dangerous that is a case of when we choose higher value of t value. Similarly, when you go for lower value of t value, the person may not have the disease but you are going to give a report saying that he has the disease, so both are dangerous. Now, you see the second category; if t is small, predict negative rarely when P of y equal to small.

More errors where we say positive but it is actually negative because we shifted the line to extreme left hand side, so that fellow is a positive but actually, it is negative, he is not having the disease, it detects all patients who are positive, whoever goes to that laboratory, they will get a report of that saying that you have the disease. So, with no preference between errors, you can select t equal to 5%.

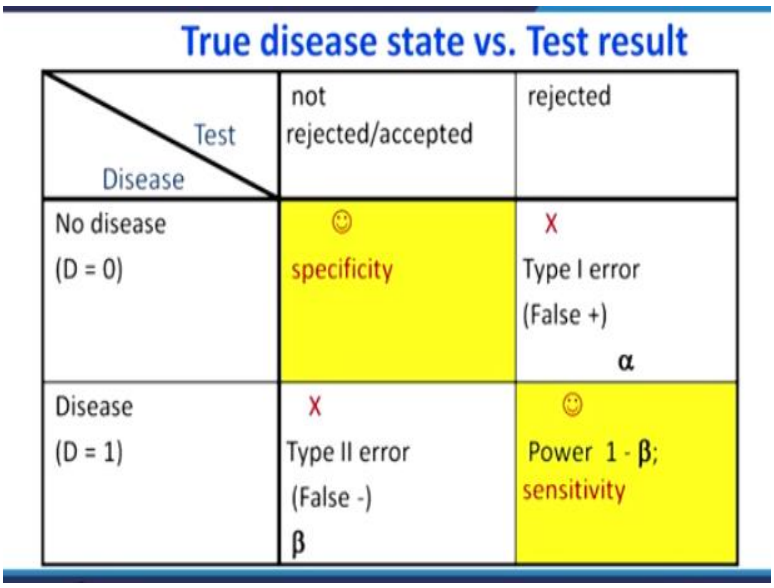
Suppose, if you are not knowing, you are not able to say the cost of that error false positive or false negative, you can keep t equal to 0.5, it predicts the more likely outcome, it is a very conservative way.

(Refer Slide Time: 23:33)



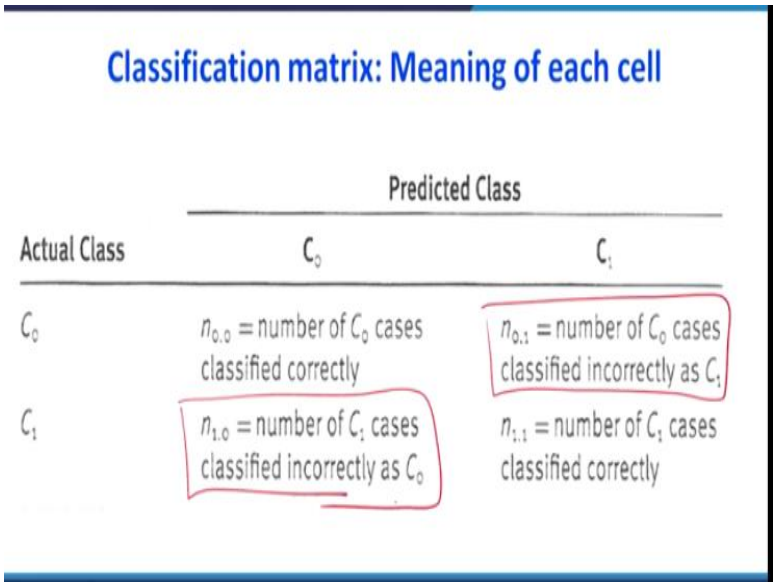
Now, I have brought this saying what is a true negative, true positive, selecting the threshold value, compare actual outcomes to predicted outcomes using confusion matrix, this also I have shown you. See that 0 0 it is a true negative when this is a false positive actually, this person is not having disease but you have given a report saying that he has a disease, this is a false negative; here false negative is this fellow actually having the disease.

(Refer Slide Time: 24:07)



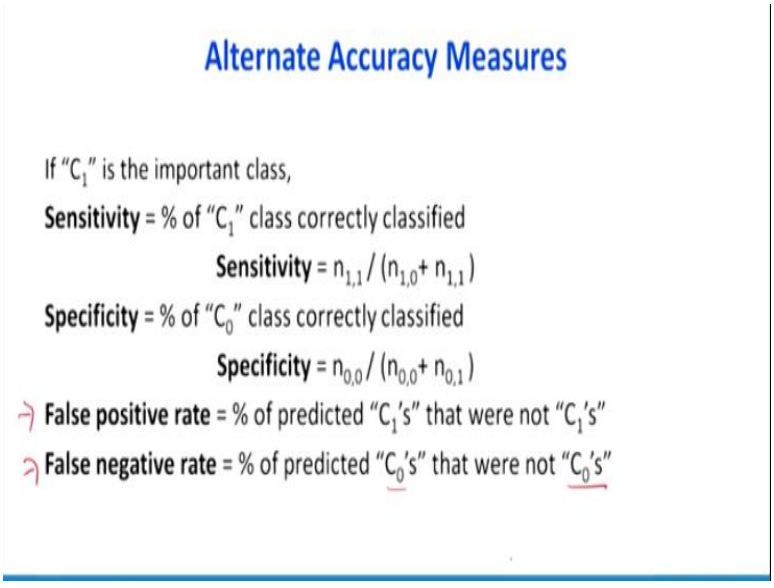
But we have given a report saying that it is negative, this says the false positive is nothing but your alpha type I error. Here these type II false negative is nothing but beta, it is your type II error, this power of test we used to say in hypothesis testing 1 – beta, it is called sensitivity. If it is actual also 0, the predicted also 0, we say it is specificity.

(Refer Slide Time: 24:34)



You see that in the term of; in the form of matrix say, C0, C0, so number of C0 cases classified correctly, you come to this diagonal, C1 and C1, n1 1 equal to number of C1 cases classified correctly, the error comes here. What it says, actually it is 0 but we have predicted as 1. Similarly, this error has come, actually it is 1, we predicted as 0, so this is the confusion matrix.

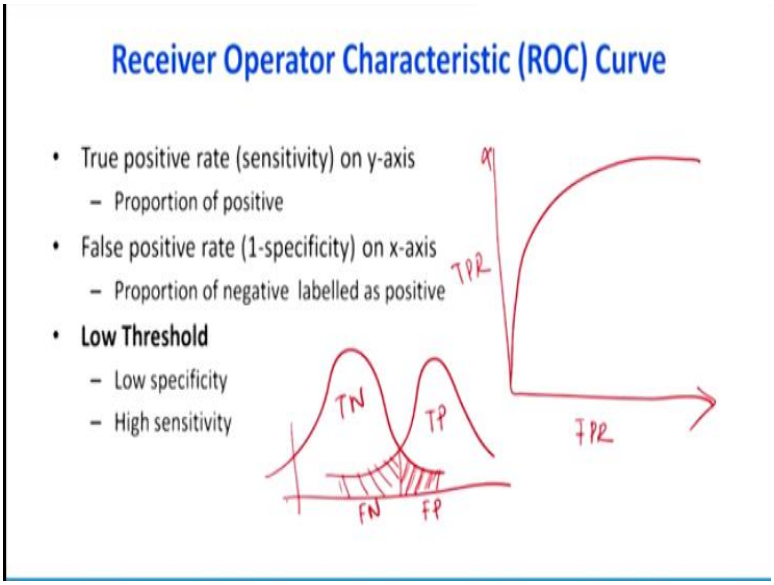
(Refer Slide Time: 25:08)



Now, let us explain this term what is sensitivity and specificity with respect to the previous table. If C1 is important class, the sensitivity equal to percentage of C1 class correctly classified, so sensitivity equal to n1 1 divided by n1 0 + n1 1. Specificity equal to percentage of C0 class correctly classified, specificity equal to n0 0 divided by n0 0 + n0 1, you can look at my previous slide; you can see that how it has come.

The, what is a false positive rate; percentage of predicted C1's that were not C1 that is a false positive rate, false negative rate is percentage of predicted C0's that were not C0 that is false positive, false negative rate because these terms we are going to use while constructing our ROC curve that is why I am defining what is a false; then we will say true positive and false positive, okay.

(Refer Slide Time: 26:23)



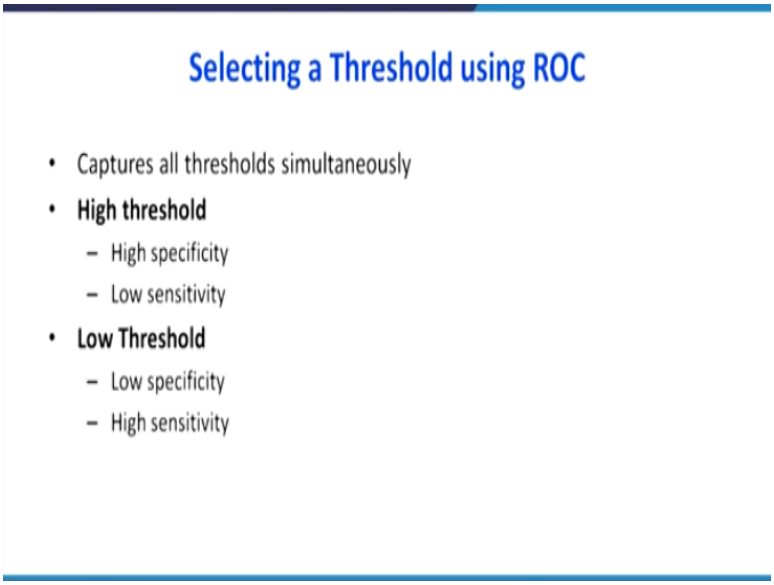
Then, what is a true positive rate; receiver operating characteristic curve, there will be, it will be go this way, in y axis, we have TPR, true positive rate. In x axis we will have 1 minus specificity that is false positive rate, so what will happen, receiver operating characteristic curve, the structure of ROC curve is in x axis, we will have true positive rate, this one, true positive rate that will be in y axis that is the proportion of positive cases.

In x axis, we are going to see false positive rate, what is a false positive rate; this false positive rate this portions which I shaded with this one, this is a false positive rate that we are going to

explain, this is 1 minus specificity, that will give your false positive rate, FPR. Now, what will happen; the curve will go this way, I will explain what will happen; when you keep very low threshold value, will have high sensitivity and low specificity.

Because low threshold in the sense, cut-off is here, there will be high sensitivity, so whoever comes to the laboratory, we will say that he has a disease, so the opposite of this is low specificity.

(Refer Slide Time: 28:44)



The another category what will happen, when you keep higher threshold, it will have high specificity, very low sensitivity. So, whoever goes there will get a report saying that he is not having the disease, so high specificity. So, there is a contradiction of keeping higher cut-off value and lower cut-off value, so we have to choose the trade-off between the both the errors that we will see in the next class.

In this lecture, we have seen how to check the quality of our regression model, there are 2 methods; one is confusion matrix, another one is ROC analysis. I have explained using confusion matrix what is a difference cell means, what is an intuitive understanding of each cell that is what is a false positive and false negative, then I have explained some theory about the ROC analysis, then I have explained what will happen when the cut-off ratio is higher what will happen, when the cut-off ratio is very low, what will happen to that.

Now, in the next class we are going to see how to choose the correct cut-off value, so that in the next class in pictorially I will explain how to choose the correct cut-off value, so that there will be a trade-off between false positive and false negative error, thank you.