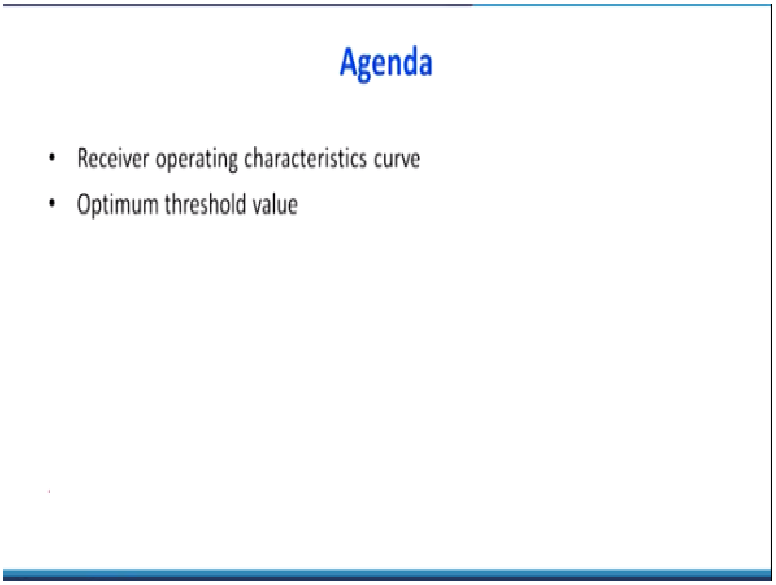
Data Analytics with Python

Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 42
Confusion Matrix and ROC - II

In this lecture, we will continue with our explanation of ROC analysis, in our previous lecture I have given you a theory about what is the confusion matrix and ROC analysis. In this lecture I will explain pictorially what is the different types of ROC curve, how that ROC curve is used to choose a correct classifying methodology that is ROC can help to predict the accuracy of our regression model.

(Refer Slide Time: 00:56)

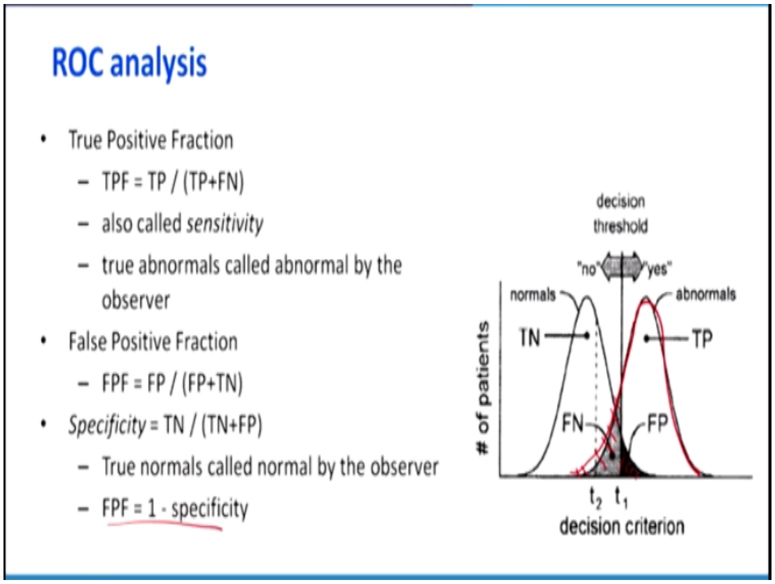


Agenda

- Receiver operating characteristics curve
- Optimum threshold value

The agenda for this lecture is we will continue with receiver operating characteristic curve, then how to choose the optimal threshold value to classify the category whether it is 1 or 0, here we will use lot of pictures give you more understanding for you.

(Refer Slide Time: 01:12)



ROC analysis

- True Positive Fraction
  - TPF = TP / (TP+FN)
  - also called *sensitivity*
  - true abnormals called abnormal by the observer
- False Positive Fraction
  - FPF = FP / (FP+TN)
- *Specificity* = TN / (TN+FP)
  - True normals called normal by the observer
  - FPF = 1 - specificity

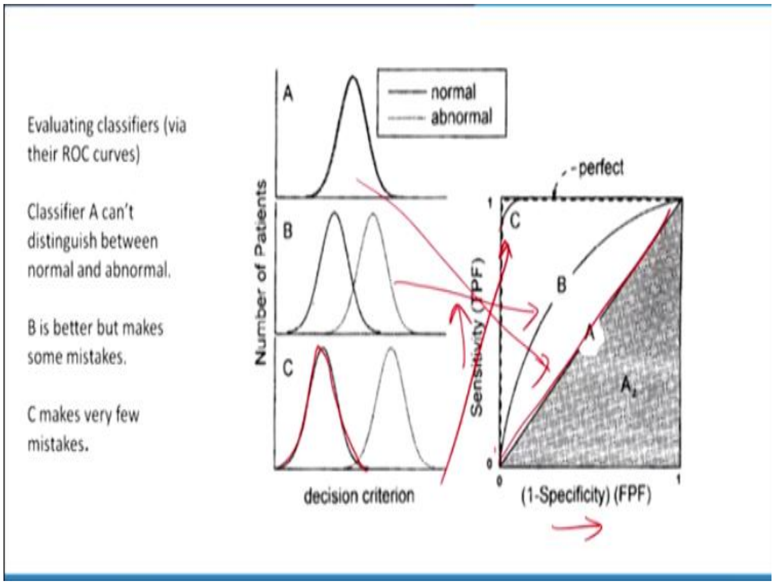
What is ROC analysis? As I told you the curve which I am showing previously, see this portion, this portion is abnormals, person having disease, so if you are predicting exactly it is a true positive, sometime this much portions actually, he is belongs to normals category, since it is lying on the abnormal side, we have given you report saying that you have the disease that is a false positive.

Now, this is the decision threshold, you see the another choice that whoever on the left hand side, we are going to say true negative, in the true negative sides, there are some people who belongs to positive side, their also lying on this side, in the negative side, so we are going to give a report false negative. What is a false negative? Even though they have the disease, we are going to say that you do not have the disease.

So, what is a true positive fraction; true positive fraction is true positive divided by your false negative, it is 1, also called sensitivity, true abnormals called abnormal by the observer, this is the right way because if they are abnormal, we are going to say that yes, they are abnormal. False positive fraction is FPF, is a false positive divide by false positive plus true negative, then specificity; true negative divided by true negative and false positive.

True normals are called normals by the observer, this false positive rate is nothing but 1 minus specificity, so what will happen; false positive rate if you write 1 minus something you will end up with false positive divided by true negative plus false positive.

(Refer Slide Time: 03:23)



Evaluating classifiers (via their ROC curves)

Classifier A can't distinguish between normal and abnormal.

B is better but makes some mistakes.

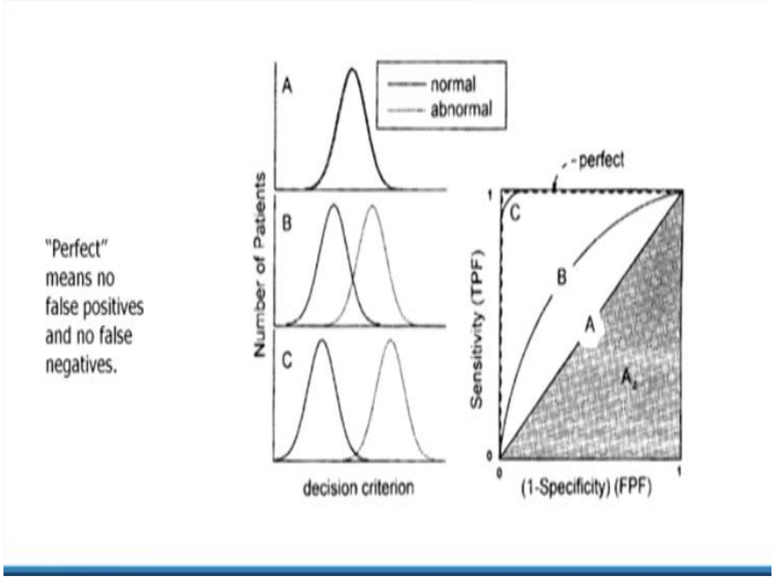
C makes very few mistakes.

The next one is the true positive fraction, TPF equal to true positive, this curve is ROC curve as I told you in y axis, we have true positive fraction, in x axis we have false positive fraction, see the range is 0 0 because it is a problem it is 0 to 1, in y range also 0 to 1 1, this curve, the C curve you see that there is a C, you see that this category, when there is a somewhat overlap between this is a true negative, true positive, so this is the situation of your so this C.

So, we can I am writing here, so this was for this curve, you see that one there are some more, more overlap, so this is the situation when the B curves, you see there is a complete overlap, this is the situation of our A curve. If it is completely separated both true positive, true negative is 2 separate curves, so you will get a perfect ROC curve. Now, look at the different conditions, evaluating classifiers via ROC curves, classifier A, this one, this line cannot distinguish between normal and abnormal.

Because it is normal and abnormal there are 2 curves which are completely overlapped, the second one B is better but makes some mistakes, this situation because there is a somewhat overlap, C makes very few mistake, it is not completely separate, so this line, the reverse L shape is perfect line. So, this graph, this picture connects between ROC curve and different types of your true positive and true negative.

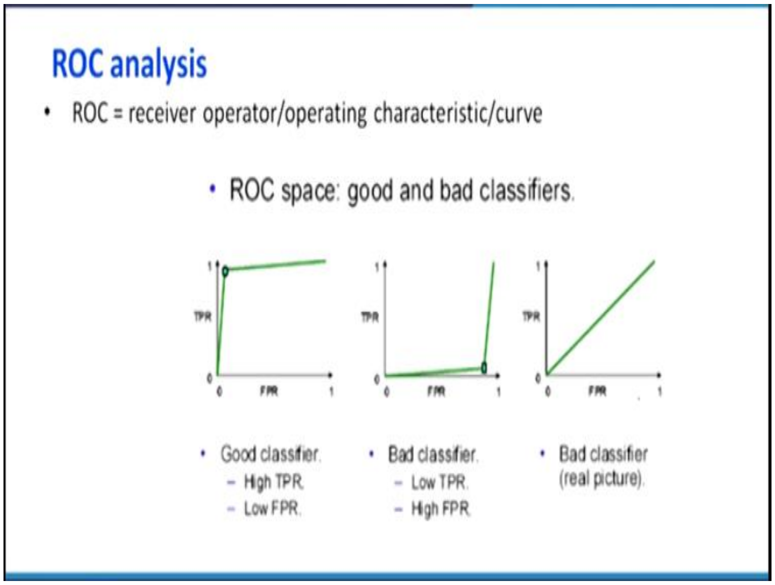
(Refer Slide Time: 05:16)



"Perfect" means no false positives and no false negatives.

There may be a perfect category, a perfect means no false positive and no false negative, so this line, no positive; no false positive and no false negative.

(Refer Slide Time: 05:32)



ROC analysis

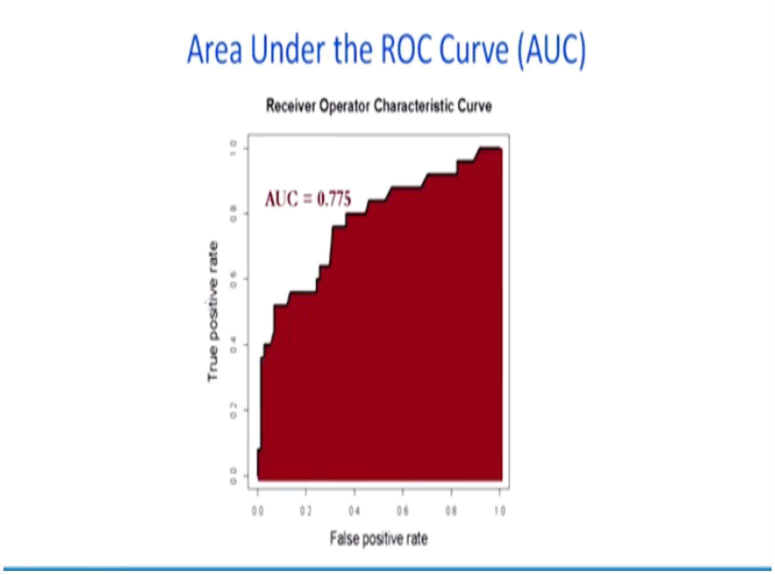
- ROC = receiver operator/operating characteristic/curve

- ROC space: good and bad classifiers.

- Good classifier.
  - High TPR.
  - Low FPR.
- Bad classifier.
  - Low TPR.
  - High FPR.
- Bad classifier (real picture).

Look at there are different situations, see ROC curve; receiver operating or OC curve, ROC space good and bad classifier, you look at this one, this is a good classifier. Why we are saying it is a good classifier? High true positive rate and low false positive rate, this one we say bad classifier because low true positive rate but high false positive rate, this one is a bad classifier because real picture because it is both are equal.

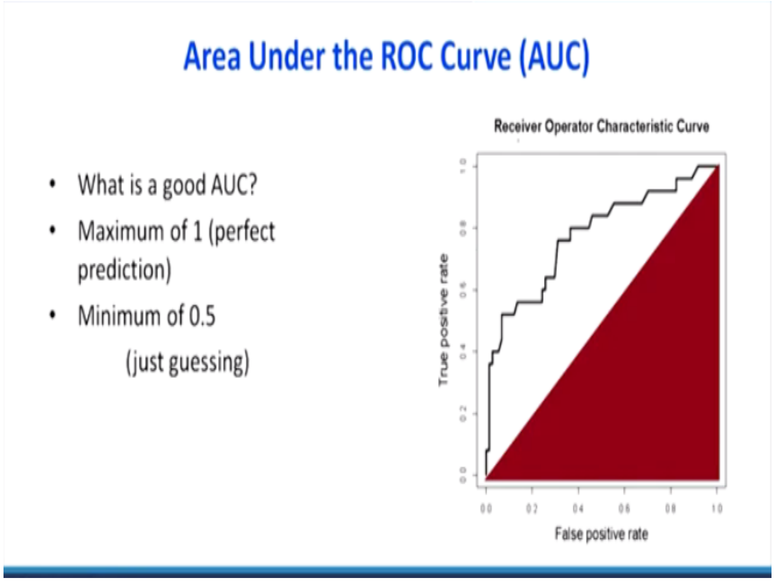
(Refer Slide Time: 06:10)



Area Under the ROC Curve (AUC)

Next there is a one more term to explain the quality of a regression model that is area under curve, AUC that is nothing but in ROC curve, the area under ROC curve it is called AUC, area under curve. For example here AUC is 0.775, this portions where the red in colour, so all other things are same, the true positive rate, false positive rate.

(Refer Slide Time: 06:36)

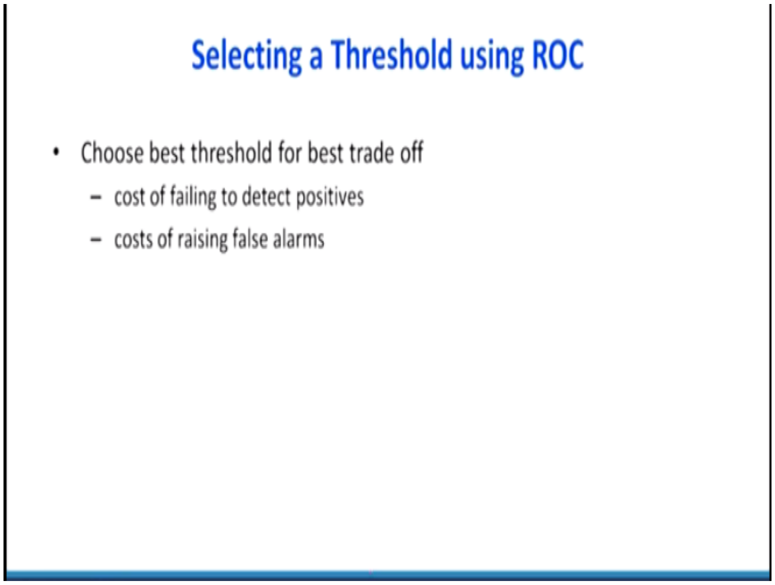


Area Under the ROC Curve (AUC)

- What is a good AUC?
- Maximum of 1 (perfect prediction)
- Minimum of 0.5 (just guessing)

What is a good AUC, area under curve, see maximum it can hold everything that is a perfect prediction, so if the perfect prediction is there, that means, all 0’s are predicted 0, all 1’s are predicted as 1, then you will get AUC, this full red colour. What is a good area under curve? The maximum value is 1 and minimum value is 0.5, it is just guessing one, so maximum it can go up to 1.

(Refer Slide Time: 07:09)

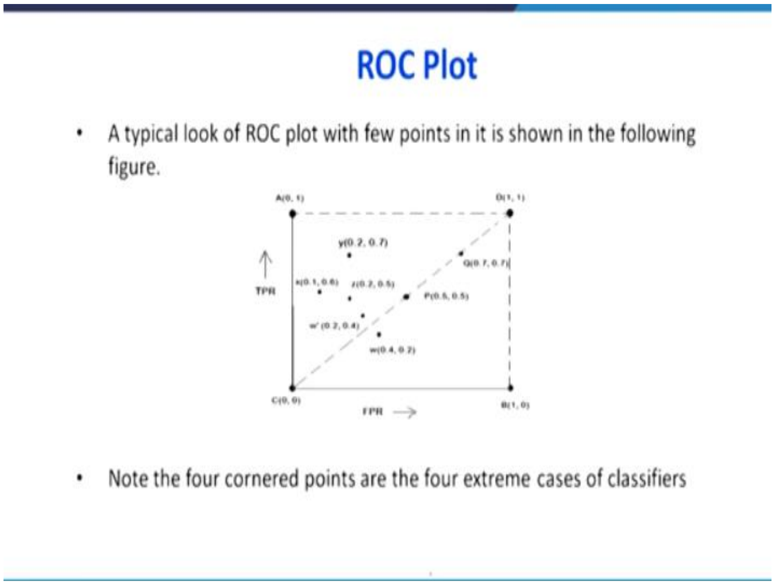


Selecting a Threshold using ROC

- Choose best threshold for best trade off
  - cost of failing to detect positives
  - costs of raising false alarms

Now, we will look for selecting a threshold using ROC curve, we have seen what is the different point of ROC curve, how to choose the threshold value that was important objective of this lecture. Choose the best threshold for best trade off, we are looking at cost of failing to detect positive and cost of raising false alarm, it is like a false positive and false negative, we have to see cost of that 2, whichever is more dangerous or more costly that should be minimised.

(Refer Slide Time: 07:42)



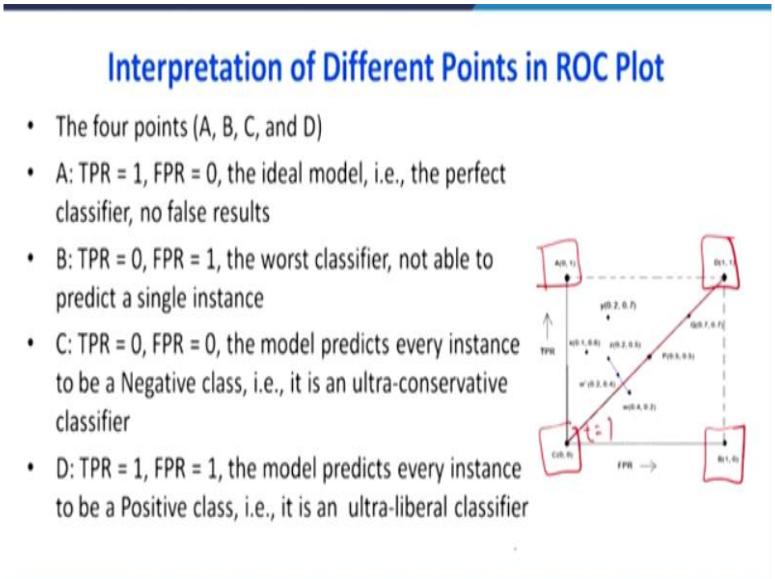
ROC Plot

- A typical look of ROC plot with few points in it is shown in the following figure.

- Note the four cornered points are the four extreme cases of classifiers

Now, we will explain ROC plot each corners, a typical ROC plot with a few point in its shown in the following figure, there are 4 point is there; A, B, C, D, note that the 4 corner points are 4 extreme case of classifier, there are different points which are above the diagonal, some points are below the diagonal, we will take each and every points I will explain what is the significance of these points and how to interpret this point.

(Refer Slide Time: 08:17)



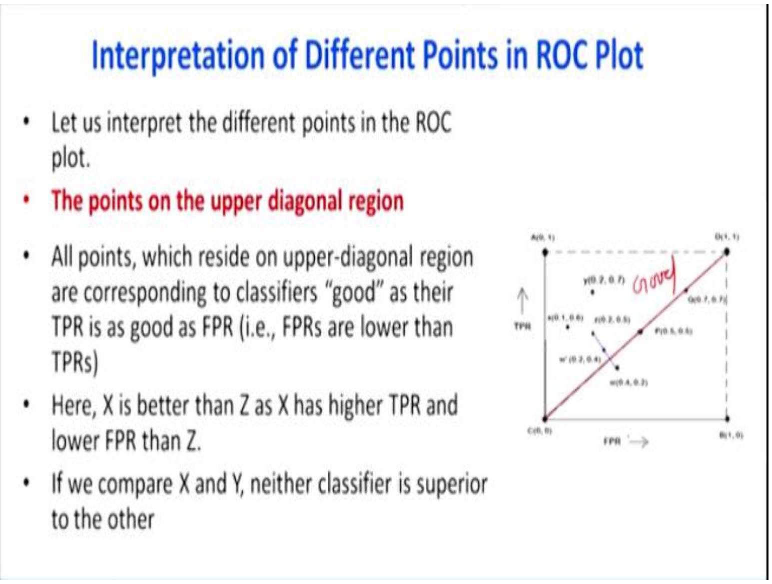
Interpretation of Different Points in ROC Plot

- The four points (A, B, C, and D)
- A: TPR = 1, FPR = 0, the ideal model, i.e., the perfect classifier, no false results
- B: TPR = 0, FPR = 1, the worst classifier, not able to predict a single instance
- C: TPR = 0, FPR = 0, the model predicts every instance to be a Negative class, i.e., it is an ultra-conservative classifier
- D: TPR = 1, FPR = 1, the model predicts every instance to be a Positive class, i.e., it is an ultra-liberal classifier

First, we will look at the point A; point A is this location, what is happening here; the true positive y value is 1, false positive is 0, this is an ideal model, the perfect classifier no false result, then we will go to the second category that is B, here the true positive rate is 0 because y value is 0 but x value is 1. The worst classifier not able to predict a single instance, then we will go for C, this situation where true positive also 0, false positive also 0.

The model predicts every instance to be negative class, it as an ultraconservative classifier, this will happen when t equal to 1. Suppose, if you keep a threshold that is very high level so everybody will be called it as negative class, the D this point is true positive rate also 1, false positive rate also 1, the model predict every instance to be positive class. When you take threshold value extreme left hand side so, whoever comes to the pathology laboratory, we will say that you have the disease, it is an ultra-liberal classifier.

(Refer Slide Time: 09:48)



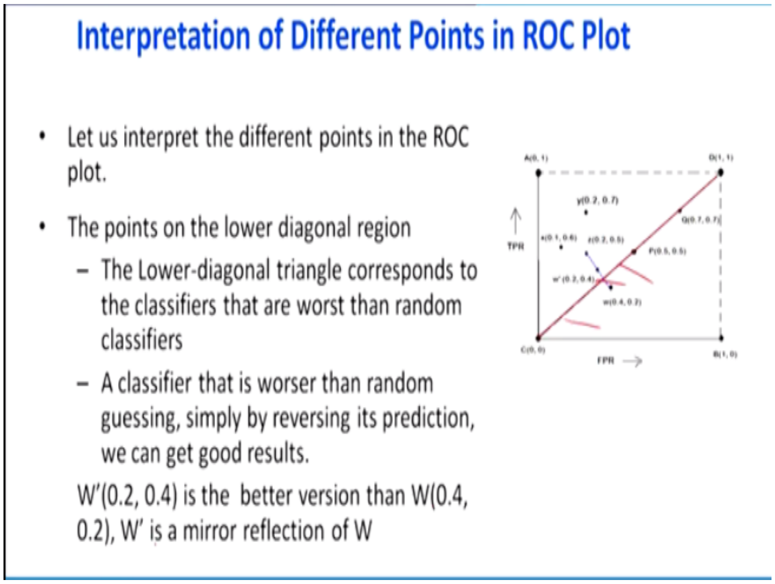
Interpretation of Different Points in ROC Plot

- Let us interpret the different points in the ROC plot.
- **The points on the upper diagonal region**
- All points, which reside on upper-diagonal region are corresponding to classifiers "good" as their TPR is as good as FPR (i.e., FPRs are lower than TPRs)
- Here, X is better than Z as X has higher TPR and lower FPR than Z.
- If we compare X and Y, neither classifier is superior to the other

The problem comes how to choose the right ROC value, now look at different points inside the ROC curve. First we will look at the points on the upper diagonal region, all points which resides on upper diagonal region or corresponding to classifiers good because this portion is the good classifier, as their true positive rate is as good as false positive rate that is false positive rate is lower than the true positive rates.

See there is one point X, when you compare X and Z, X is better than Z because X has higher true positive rate and lower false positive rate than Z, when compare to X and Z, X is better. If you compare X and Y see that neither classifier is superior because there is a trade-off between TPR and FPR, if the TPR is increasing, FPR also increasing.

(Refer Slide Time: 10:55)



Interpretation of Different Points in ROC Plot

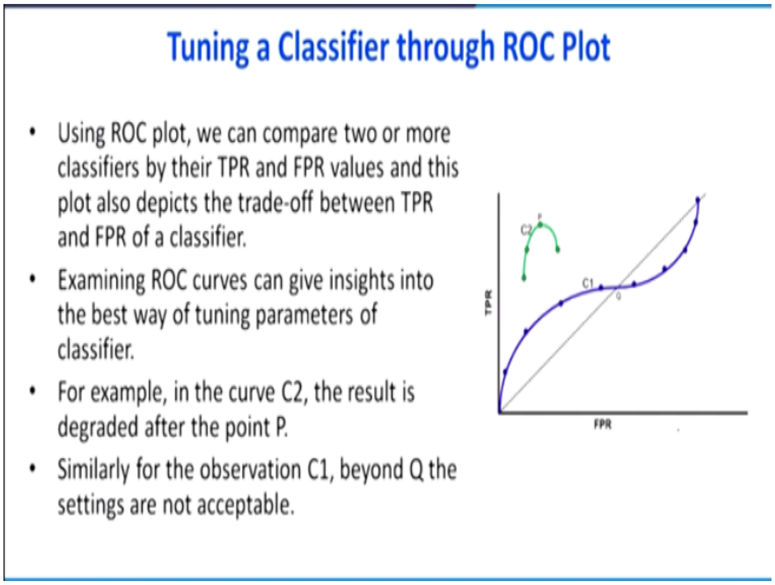
- Let us interpret the different points in the ROC plot.
- The points on the lower diagonal region
  - The Lower-diagonal triangle corresponds to the classifiers that are worst than random classifiers
  - A classifier that is worser than random guessing, simply by reversing its prediction, we can get good results.

  W'(0.2, 0.4) is the better version than W(0.4, 0.2), W' is a mirror reflection of W

Now, let us interpret the different point ROC curve, first we will see the points on the lower diagonal region, previously I have explained the upper diagonal region, now we will look at the points in the lower diagonal region. The lower diagonal triangle corresponds to the classifier that are worse than the random classifier because this side it is not good because it is a high false positive rate.

A classifier that is worser than the random guessing simply by reversing its prediction, suppose look at the 2 point W dash and W, W dash is 0.2, 0.4 is better version than the W 0.4 and 0.2 because W dash is the mirror reflection of W.

(Refer Slide Time: 11:41)



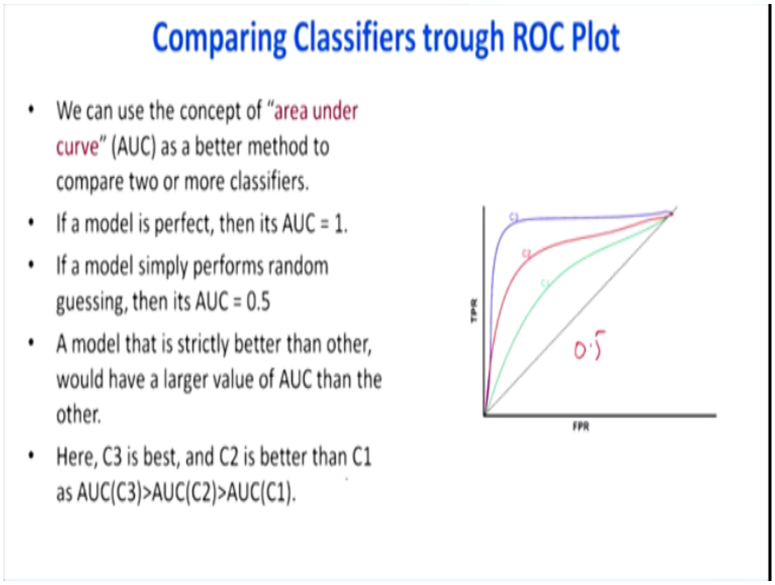
Tuning a Classifier through ROC Plot

- Using ROC plot, we can compare two or more classifiers by their TPR and FPR values and this plot also depicts the trade-off between TPR and FPR of a classifier.
- Examining ROC curves can give insights into the best way of tuning parameters of classifier.
- For example, in the curve C2, the result is degraded after the point P.
- Similarly for the observation C1, beyond Q the settings are not acceptable.

Now, tuning a classifier through ROC plot, see that I have 2 category of ROC plot, let us see which is better, why. Using ROC plot, we can compare 2 or more classifier by their TPR that is a true positive rate and false positive rate values and this plot also depicts the trade-off between true positive rate and false positive rate of a classifier. Examining ROC curves can give insight into the best way of tuning parameter of classifier.

For example, in this curve C2, the result is degraded after the point B, you see that this C2, what is happening; true positive rate is increasing after point B, what is happening; true positive rate is decreasing but there is no much decrease on false positive rate, so beyond this point P, it is not giving good classification. Similarly, for the observation C1, beyond Q, the setting are not acceptable because there is a comparatively lower false positive rate when compared to true positive rate.

(Refer Slide Time: 12:54)



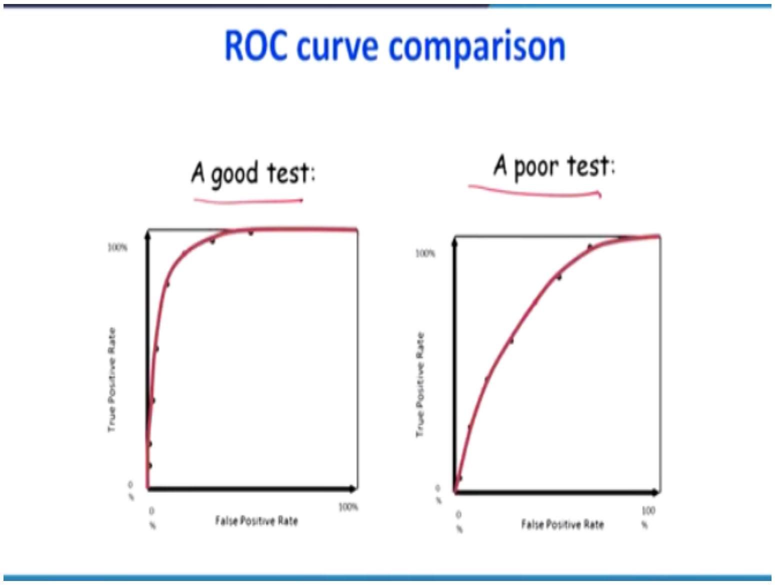
Comparing Classifiers trough ROC Plot

- We can use the concept of "area under curve" (AUC) as a better method to compare two or more classifiers.
- If a model is perfect, then its AUC = 1.
- If a model simply performs random guessing, then its AUC = 0.5
- A model that is strictly better than other, would have a larger value of AUC than the other.
- Here, C3 is best, and C2 is better than C1 as AUC(C3)>AUC(C2)>AUC(C1).

Now, there are different classifying comparing different classifiers through ROC plot, when you look at this picture, see that there are C1, C2, C3 we can use the concept of area under curve as a better method to compare 2 or classifier, we can get different classifier by getting different threshold value. If a model is perfect, then the AUC is 1 which I have seen; which I have explained.

If a model simply performs a random guessing then the AUC is 0.5, so this area, a model that is strictly better than other would have larger value of AUC, area under curve than the other. So, out of these 3, the C3 is having higher area under curve, so that model and that corresponding threshold value is better to classify which is good or which is 1 or which is 0, when compared to C2, C1. Here, the C3 is the best, C2 is better than C1 as AUC, area under curve C3 is greater, then AUC , area under curve C2 greater than area under curve C1.

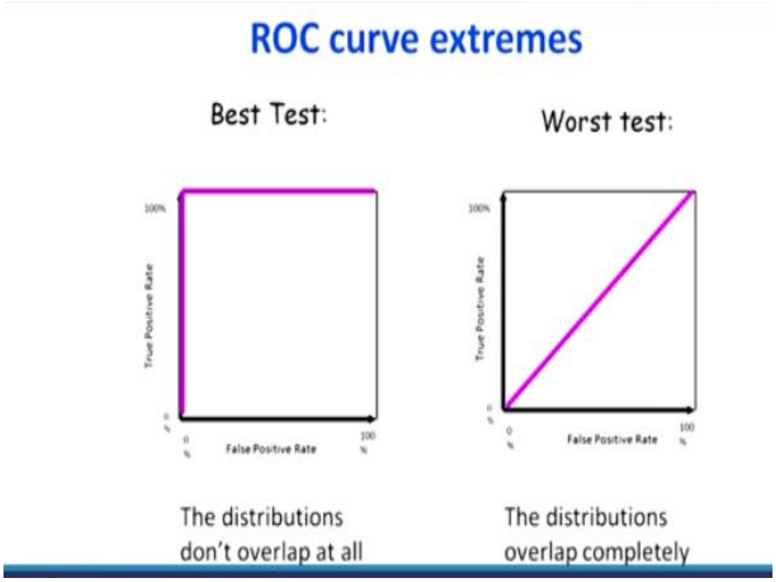
(Refer Slide Time: 14:17)



ROC curve comparison

Now, let us look at our extreme cases of this ROC curve, this was our typical ROC curve, see that how to compare 2 type of ROC curve, you see that it is closer, the area under curve is somewhat nearer to 1, so it is a good test. When you look at this one, area under curve of that ROC curve is lesser when compare to this, so it is a poor test. The left side one is the best test sorry, good test.

(Refer Slide Time: 14:55)



ROC curve extremes

Best Test:

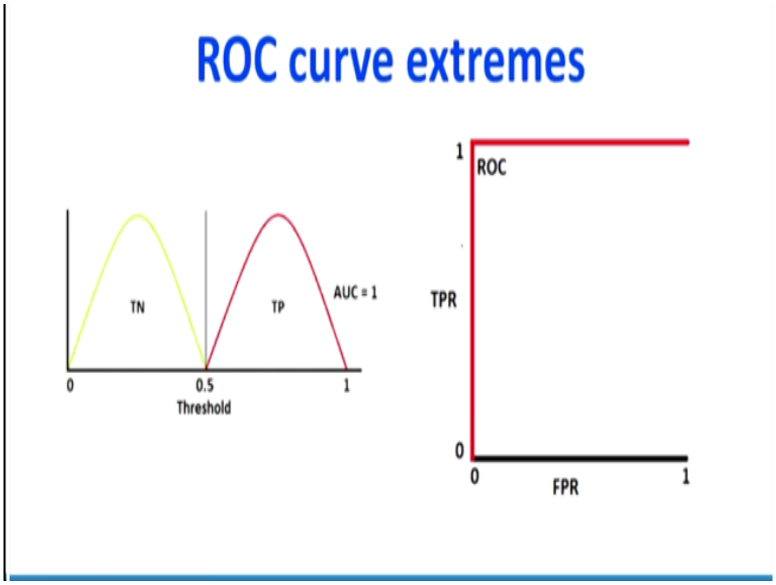
Worst test:

The distributions don't overlap at all

The distributions overlap completely

Now, we will see extreme cases, when this extreme cases, see that the 2 distributions do not overlap at all, in the previous lecture I have shown you 2 cases, one is true negative and positive, this side true negative, true positive, then false positive, false negative. If the 2 lines are; 2 distributions are not overlapping, we will get very ideal case that is best test. The distributions will overlap completely, then you will get a this kind of diagonal, this is a worst test.

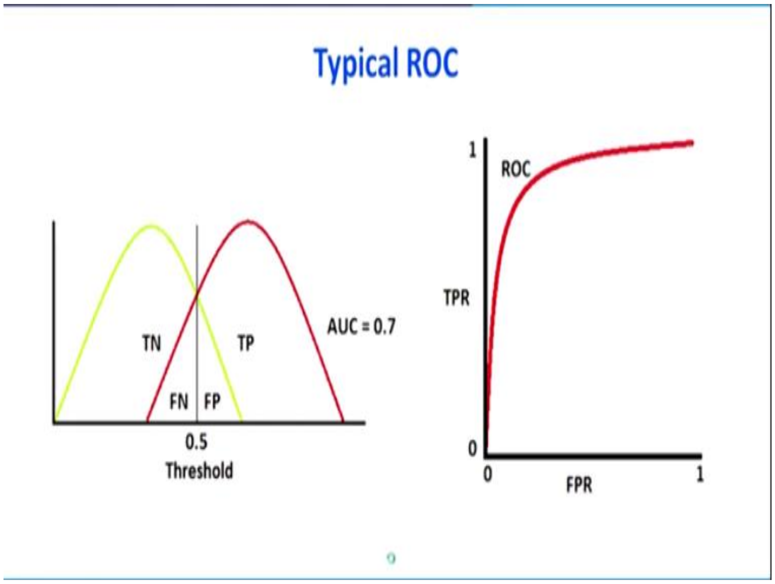
(Refer Slide Time: 15:32)



ROC curve extremes

You see that this case, true negative true positive, there is no overlap at all, so you will get this kind of ROC curve.

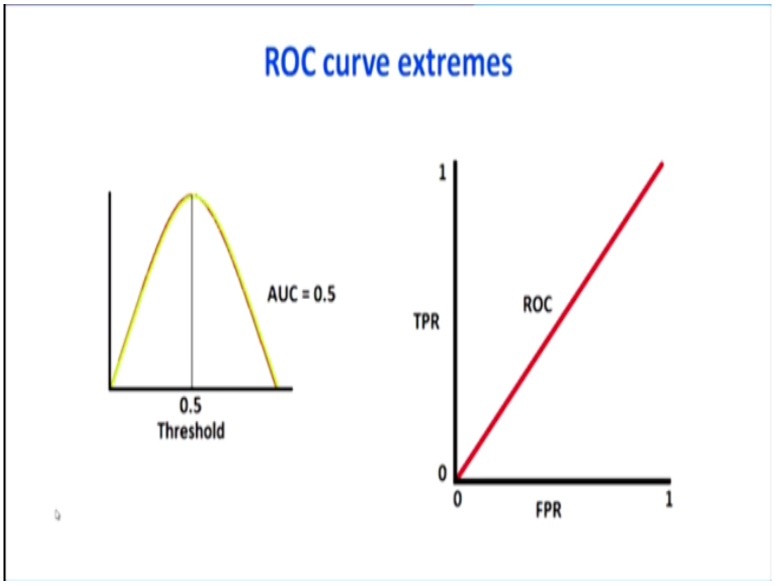
(Refer Slide Time: 15:43)



Typical ROC

You see that there are somewhat some overlap is there, so you will get a typical ROC curve because there is area under curve, previously it was 1, now it is area under curve is 0.7.

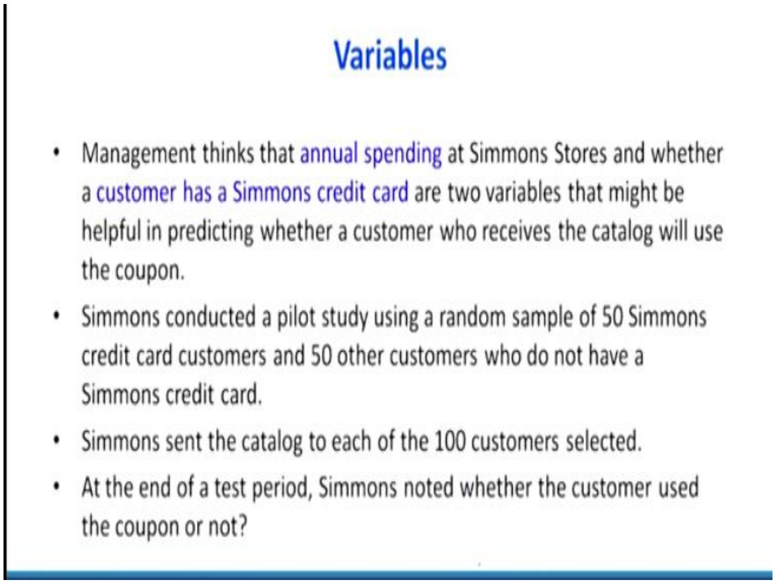
(Refer Slide Time: 15:57)



ROC curve extremes

Now, you see this case the area, both the distributions are completely overlapping that whenever it is overlapping, area under curve is 0.5, so corresponding ROC curve will look like this, so far we have seen and we have understood the concept of confusion matrix and ROC curve, now I am going to take one example, that example already I have discussed in my previous lectures. With the help of that example I am going to tell you how to choose the correct threshold value to classify whether it is belongs to category 1 or category 2.

(Refer Slide Time: 16:38)

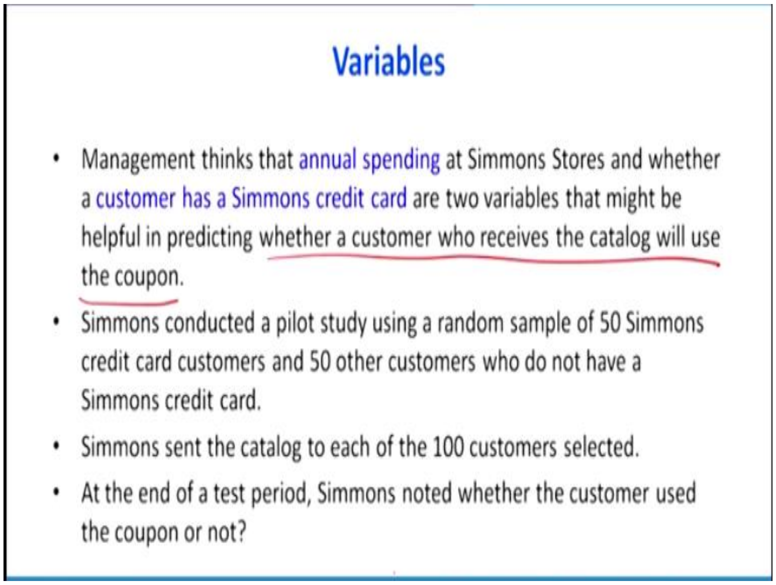


Variables

- Management thinks that annual spending at Simmons Stores and whether a customer has a Simmons credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon.
- Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card.
- Simmons sent the catalog to each of the 100 customers selected.
- At the end of a test period, Simmons noted whether the customer used the coupon or not?

The example is this book; this example is taken from this book statistics for business and economics from David Anderson Sweeney and Williams. The example is; let us consider an application of logistic regression involving direct mail promotion being used by Simmons Stores. Simmons owns and operates a national chain of women's apparel stores, 5000 copies of an expensive 4 colour sales catalog have been printed and each catalog includes a coupon that provides 50 dollar discount on purchase of 200 dollar or more. The catalog are expensive and Simmons would like to send them to only those customers who have the highest probability of using the coupon.

(Refer Slide Time: 17:28)

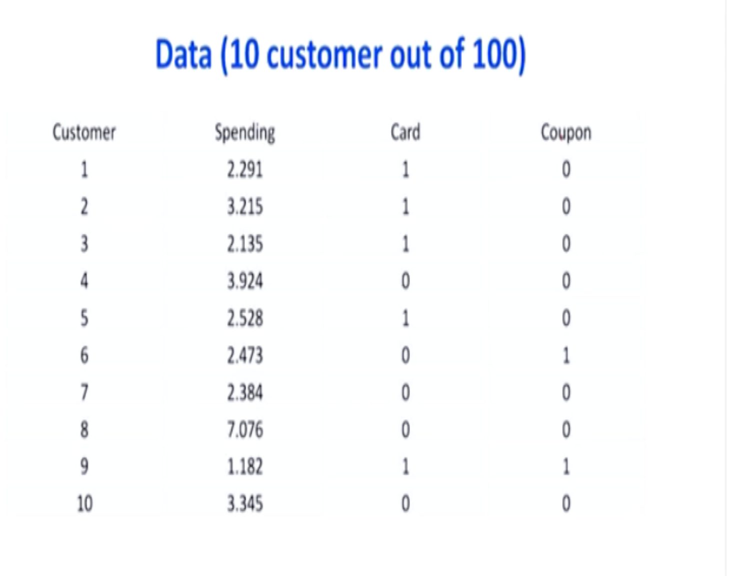


Variables

- Management thinks that annual spending at Simmons Stores and whether a customer has a Simmons credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon.
- Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card.
- Simmons sent the catalog to each of the 100 customers selected.
- At the end of a test period, Simmons noted whether the customer used the coupon or not?

So, what are the variables which are involved in this problem is; one is annual spending, another one is whether the customer has Simmons credit card or not. What we are going to predict whether a customer who receives the catalog will use the coupon or not, Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 customers who do not have Simmons credit card. Simmons sent the catalog to each of the 100 customer selected, at the end of the test period Simmons noted whether the customers used the coupon or not, this is the problem.

(Refer Slide Time: 18:11)

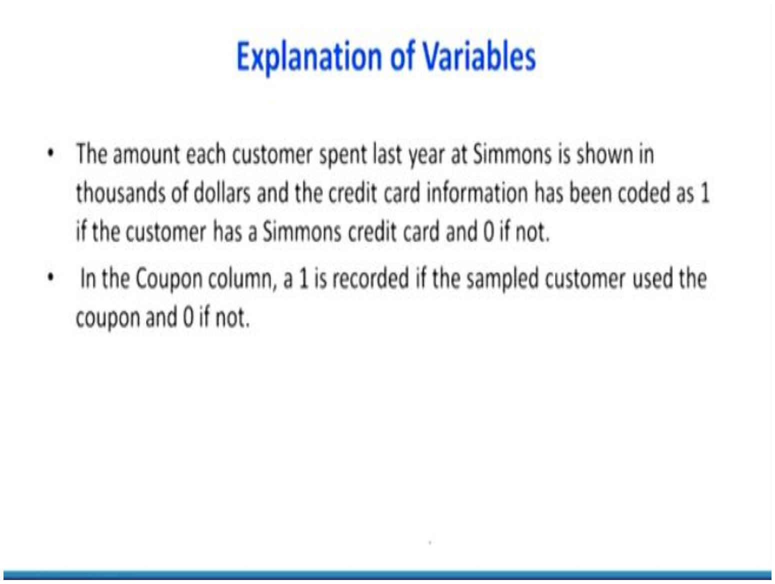


Data (10 customer out of 100)

Customer	Spending	Card	Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

For that problem this is the dataset, spending how much they spend in the last month that is our one of the independent variable. Possession of Simmons credit card, if it is 1 he has, 0 does not have that is the another independent variable. The coupon; whether he use the coupon or not that is our dependent variable.

(Refer Slide Time: 18:33)

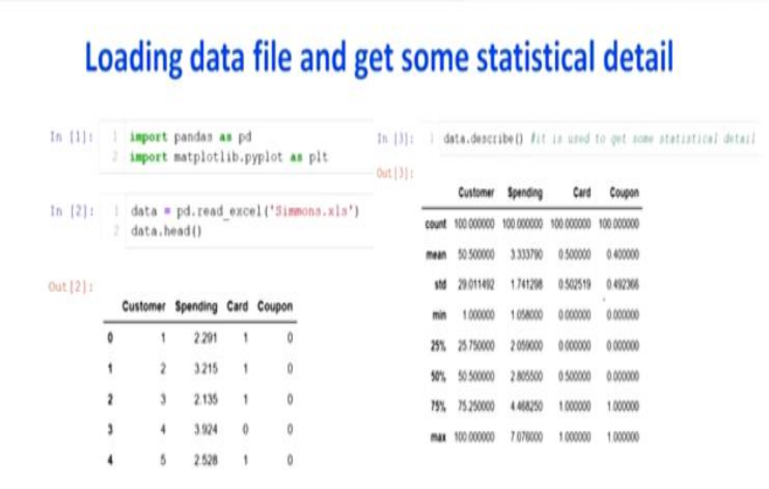


Explanation of Variables

- The amount each customer spent last year at Simmons is shown in thousands of dollars and the credit card information has been coded as 1 if the customer has a Simmons credit card and 0 if not.
- In the Coupon column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

You see that dependent variable is 2 category 0 or 1, so it is a; then we have to go for logistical regression. The amount of each customer spent last year at Simmons is shown in the 1000's of dollars and the credit card information has been coded as 1, if the customer has the Simmons credit card 0, if not. In the coupon column, 1 is recorded if the sampled customers used the coupon and 0 if not.

(Refer Slide Time: 19:04)



Loading data file and get some statistical detail

```
In [1]:  import pandas as pd
         import matplotlib.pyplot as plt
```

```
In [2]:  data = pd.read_excel('Simmons.xls')
         data.head()
```

Out[2]:

	Customer	Spending	Card	Coupon
0	1	2.291	1	0
1	2	3.215	1	0
2	3	2.135	1	0
3	4	3.924	0	0
4	5	2.528	1	0

```
In [3]:  data.describe() #it is used to get some statistical detail
```

Out[3]:

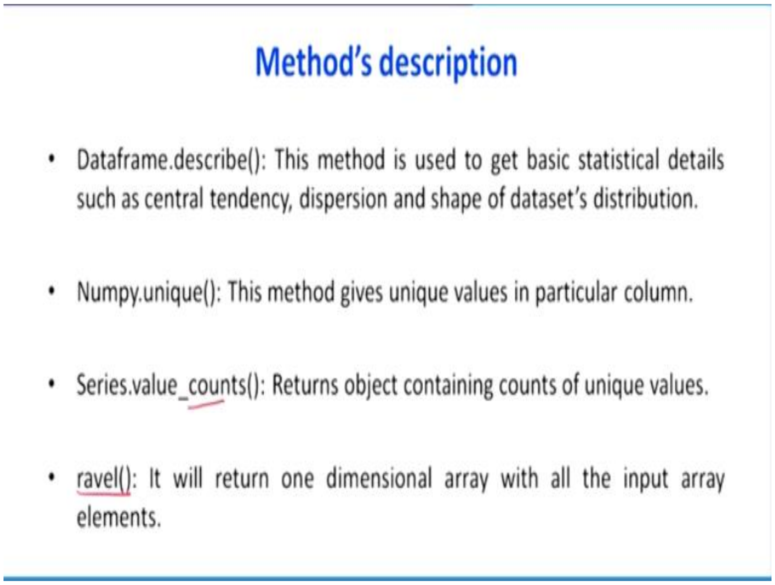
	Customer	Spending	Card	Coupon
count	100.000000	100.000000	100.000000	100.000000
mean	50.500000	3.333790	0.500000	0.400000
std	29.011492	1.741298	0.502519	0.492366
min	1.000000	1.058000	0.000000	0.000000
25%	25.750000	2.059000	0.000000	0.000000
50%	50.500000	2.805500	0.500000	0.000000
75%	75.250000	4.468250	1.000000	1.000000
max	100.000000	7.076000	1.000000	1.000000

So, we have imported the data, for import we have imported necessary libraries, import pandas as pd, import matplotlib dot pyplot as plt, then the dataset is Simmons dot xls, I am going to show, run this Python code and I am going to explain further. I brought the screenshot of my

Python output, so data dot head we came to know there is a spending is one independent variable, card and coupon.

The first one is a data dot describe, that is to get an idea about the details of each variables, customer, there is no meaning for this one, for spending you see that there are 100 values is there, the mean is 3.3, standard deviation is this one, since card is the categorical variable, there is no meaning, for the mean there is no meaning for standard deviation, so it is not applicable. The coupon also, it is a categorical variable, in the categorical variable there are 100 values is there. Here also, there is no meaning for mean and standard deviation because you cannot do any arithmetic operation when there is a nominal or categorical variable.

(Refer Slide Time: 20:17)

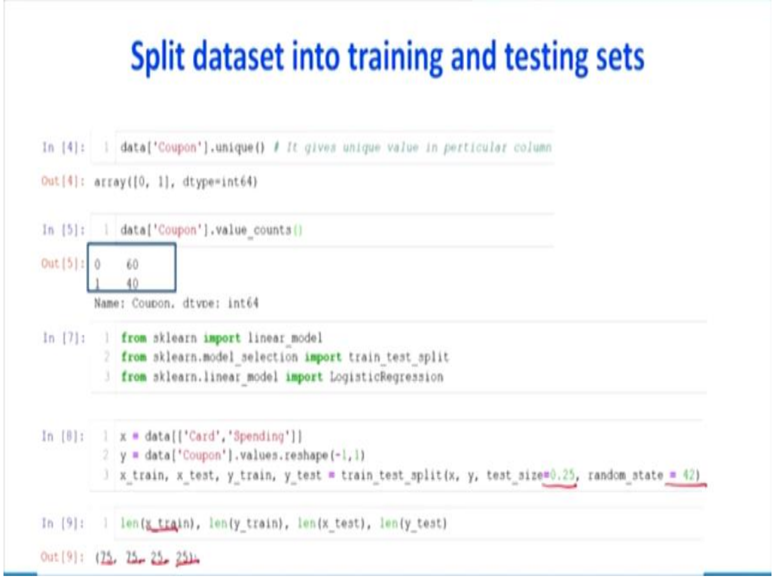


Method's description

- Dataframe.describe(): This method is used to get basic statistical details such as central tendency, dispersion and shape of dataset's distribution.
- Numpy.unique(): This method gives unique values in particular column.
- Series.value_counts(): Returns object containing counts of unique values.
- ravel(): It will return one dimensional array with all the input array elements.

We are going to use different inbuilt functions, so if you say, Dataframe dot describe that function is used to get the basic statistical details such as central tendency, dispersion and shape of the datasets distribution. If you use Numpy dot unique, this method gives the unique value in a particular column, they count this option; Series dot value underscore counts return object containing count of unique values.

(Refer Slide Time: 21:01)



Split dataset into training and testing sets

```
In [4]:  data['Coupon'].unique() # It gives unique value in perticular column
Out[4]: array([0, 1], dtype=int64)

In [5]:  data['Coupon'].value_counts()
Out[5]: 0    60
        1    40
        Name: Coupon, dtype: int64

In [7]:  from sklearn import linear_model
         from sklearn.model_selection import train_test_split
         from sklearn.linear_model import LogisticRegression

In [8]:  x = data[['Card','Spending']]
         y = data['Coupon'].values.reshape(-1,1)
         x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state = 42)

In [9]:  len(x_train), len(y_train), len(x_test), len(y_test)
Out[9]: (75, 75, 25, 25)
```

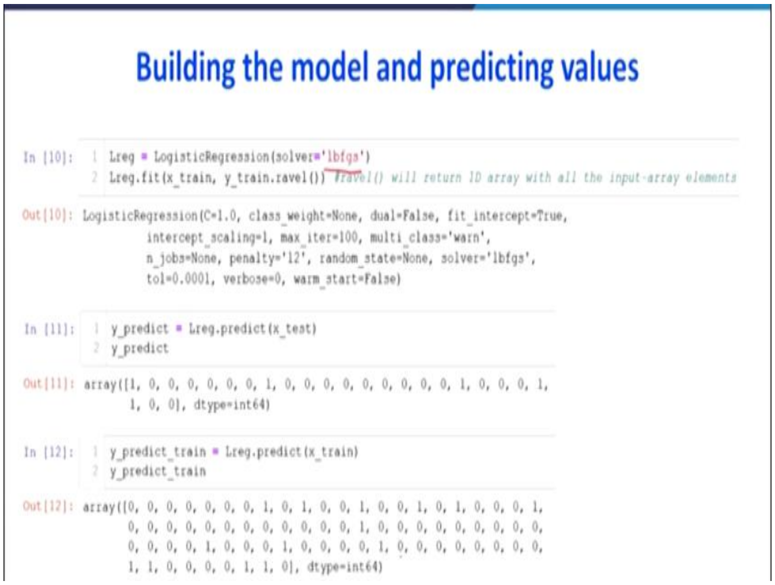
This ravel; it will return one dimensional array with all the input array elements, so I use that inbuilt function for example, data for the coupon column unique, so it is give 0, 1 because that means we can come to know there are 2 category in the coupon column, one is 0, another one is 1. Then how many when you say dot value underscore counts, say there are 60 0 40 1, 60 people did not use the coupon, 40 people have use the coupon.

Then for running logistic regression and they split the data into 2 categories; some data for training and building the model, after the model is built, we will use the test data to verify our built model from sklearn import linear underscore model, from sklearn dot model underscore import train train underscore test underscore split, from sklearn dot linear underscore model import logistic regression.

So, the x value equal to data; card and spending; independent variable, y value is coupon, okay, first we are going to split x underscore train, x underscore test that is x value after splitting we are going to call it this way then, y train y test equal to train underscore test underscore split x, y test underscore size 0.25, you can take any value, thus I have set some value, so that we can repeat the code again, you may get the same kind of output.

So, you see that I wanted to say see for the in the training data set, there is 75 data set for x and 75 data set for y. For testing data set, there are 25 data set for x, 25 data set for y, this is for testing purpose.

(Refer Slide Time: 22:50)



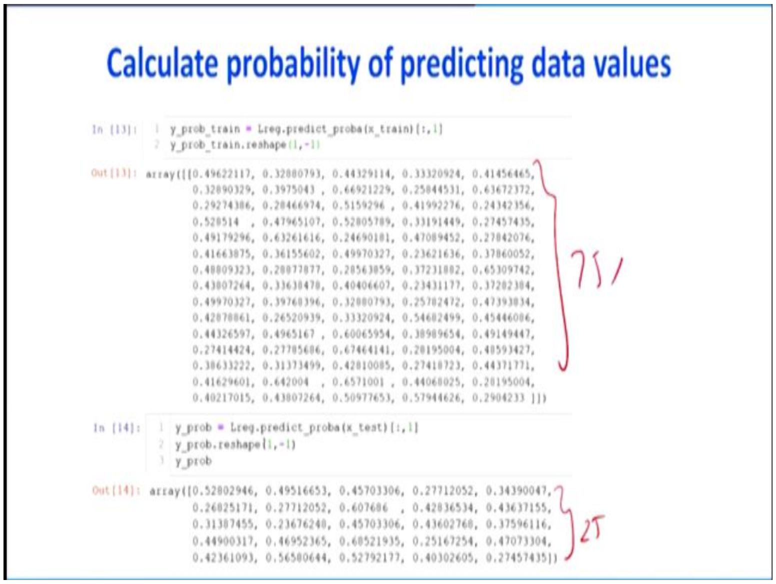
Building the model and predicting values

```
In [10]:  Lreg = LogisticRegression(solver='lbfgs')
          Lreg.fit(x_train, y_train.ravel()) #ravel() will return 1D array with all the input-array elements

Out[10]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                    intercept_scaling=1, max_iter=100, multi_class='warn',
                    n_jobs=None, penalty='l2', random_state=None, solver='lbfgs',
                    tol=0.0001, verbose=0, warm_start=False)

In [11]:  y_predict = Lreg.predict(x_test)
          y_predict

Out[11]: array([1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,
                1, 0, 0], dtype=int64)

In [12]:  y_predict_train = Lreg.predict(x_train)
          y_predict_train

Out[12]: array([0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1,
                0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
                0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
                1, 1, 0, 0, 0, 0, 1, 1, 0], dtype=int64)
```

First, we will use logistic model and will after building the model, we will predict the values, we are going to use in the logistic regression, lbfgs, there are different method I will show you there, when I am running the code, this is one method for constructing the logistic regression model, logisticregression dot fit x train y train dot ravel, we will written one-dimensional array with all input array elements, so we are getting this output.

Now, we will predict the y value by using the test data set, what I have done; we have constructed the model, then we will use a test dataset to predict the y value, this was our predicted y value, then we will predict y value by using the training dataset because training we have 75 data set, for testing there are 25 dataset, so this was after predicted the regression model using training dataset.

(Refer Slide Time: 23:52)

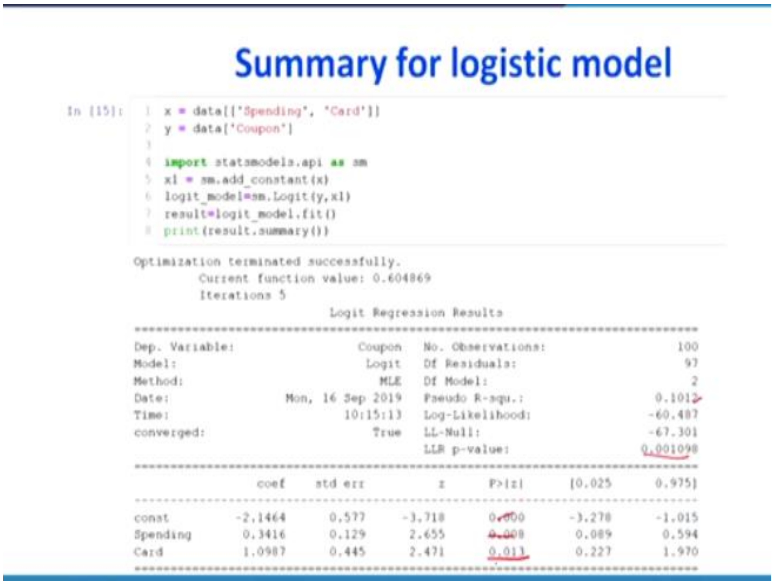


Calculate probability of predicting data values

```
In [13]:  y_prob_train = Lreg.predict_proba(x_train)[:,1]
          y_prob_train.reshape(1,-1)

Out[13]: array([[0.49622117, 0.32880793, 0.44329114, 0.33320924, 0.41456465,
                 0.32890329, 0.3975043 , 0.66921229, 0.25844531, 0.63672372,
                 0.29274386, 0.28466974, 0.5159296 , 0.41992276, 0.24342356,
                 0.528514  , 0.47965107, 0.52805789, 0.33191449, 0.27457435,
                 0.49179296, 0.63261616, 0.24690181, 0.47089452, 0.27842076,
                 0.41663875, 0.36155602, 0.49970327, 0.23621636, 0.37860052,
                 0.48809323, 0.28877877, 0.28563859, 0.37231882, 0.65309742,
                 0.43807264, 0.33638478, 0.40406607, 0.23431177, 0.37282384,
                 0.49970327, 0.39768396, 0.32880793, 0.25782472, 0.47393834,
                 0.42878861, 0.26520939, 0.33320924, 0.54682499, 0.45446086,
                 0.44326597, 0.4965167 , 0.60065954, 0.38989654, 0.49149447,
                 0.27414424, 0.27785686, 0.67464141, 0.28195004, 0.48593427,
                 0.38633222, 0.31373499, 0.42810085, 0.27418723, 0.44371771,
                 0.41629601, 0.642004  , 0.6571001 , 0.44068025, 0.28195004,
                 0.40217015, 0.43807264, 0.50977653, 0.57944626, 0.2904233 ]])

In [14]:  y_prob = Lreg.predict_proba(x_test)[:,1]
          y_prob.reshape(1,-1)
          y_prob

Out[14]: array([0.52802946, 0.49516653, 0.45703306, 0.27712052, 0.34390047,
                0.26825171, 0.27712052, 0.607686  , 0.42836534, 0.43637155,
                0.31387455, 0.23676248, 0.45703306, 0.43602768, 0.37596116,
                0.44900317, 0.46952365, 0.68521935, 0.25167254, 0.47073304,
                0.42361093, 0.56580644, 0.52792177, 0.40302605, 0.27457435])
```

Next one; y underscore probability underscore train, here see that this is the probability value which dataset for training dataset, so there are 75 dataset, we have got the probability of all the 75 dataset. Our problem is going to be there, what should be our cut-off value here, to say this below this category is called 0, above that category is 1, so y underscore probability here also we will find out for using test dataset, we will predict the probability, this is there going to be 25 dataset, here going to be 75 dataset.

(Refer Slide Time: 24:34)

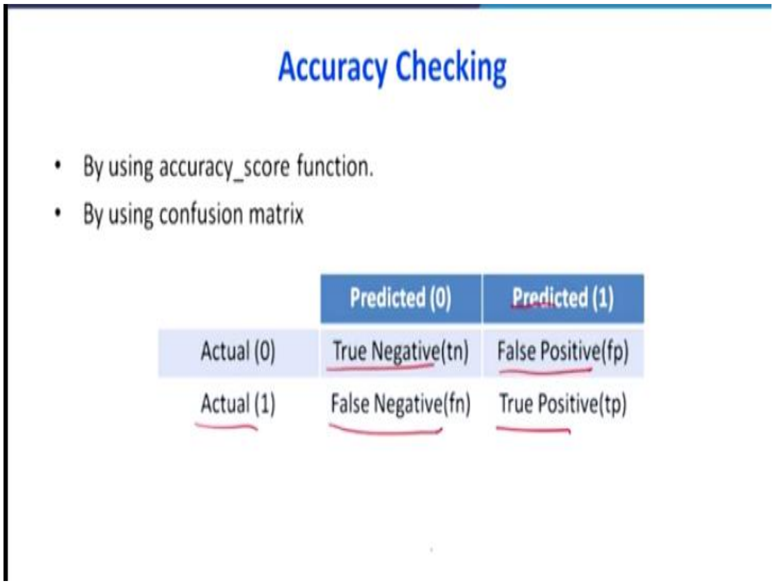


Summary for logistic model

```
In [15]:  x = data[['Spending', 'Card']]
          y = data['Coupon']

          import statsmodels.api as sm
          x1 = sm.add_constant(x)
          logit_model=sm.Logit(y,x1)
          result=logit_model.fit()
          print(result.summary())

Optimization terminated successfully.
         Current function value: 0.604869
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:                 Coupon   No. Observations:                  100
Model:                          Logit   Df Residuals:                       97
Method:                           MLE   Df Model:                            2
Date:                Mon, 16 Sep 2019   Pseudo R-squ.:                  0.1012
Time:                        10:15:13   Log-Likelihood:                -60.487
converged:                       True   LL-Null:                       -67.301
                                        LLR p-value:                  0.001098
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.1464      0.577     -3.718      0.000      -3.278      -1.015
Spending       0.3416      0.129      2.655      0.008       0.089       0.594
Card           1.0987      0.445      2.471      0.013       0.227       1.970
==============================================================================
```

This is that the; our task is what should be the cut-off value or threshold value, first we will construct a regression equation, logistic regression x is x data y data, after importing the model we got this one, constant is - 2.1464, spending independent variable is 0.3416, card 1.089, when we look at the pseudo R square, it is good, then the P value also good, the overall model is good, even if you look at the Wald test, so this P value also all are less than 0.05, this model is good.

(Refer Slide Time: 25:20)



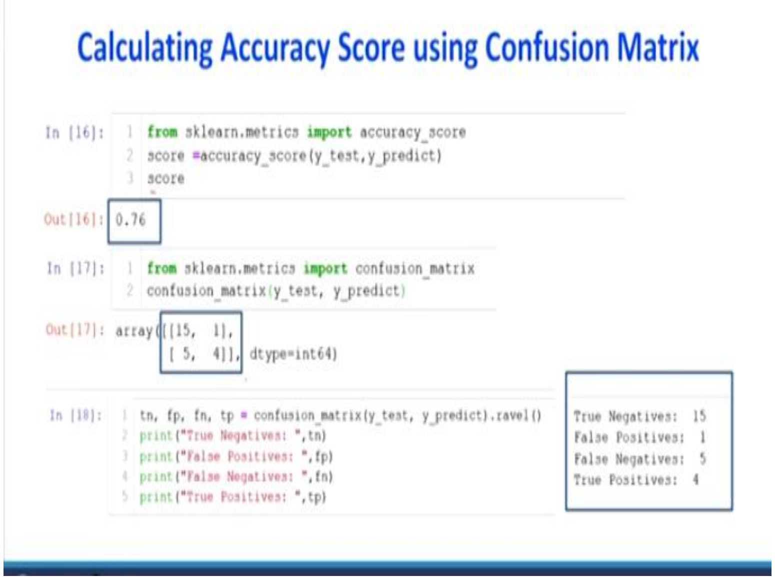
Accuracy Checking

- By using accuracy_score function.
- By using confusion matrix

	Predicted (0)	Predicted (1)
Actual (0)	True Negative(tn)	False Positive(fp)
Actual (1)	False Negative(fn)	True Positive(tp)

The point is how to set the threshold value, so after getting this model, we have predicted some portions are 1, so first we will go for checking the accuracy. There are 4 possibility; the actual is 0, predicted is 0, it is a true negative, actual is 1, predicted is 1, it is a true positive, actual is 0, the predicted is 1, it is a false positive, actual is 1, predicted is 0 that is a false negative.

(Refer Slide Time: 25:44)



Calculating Accuracy Score using Confusion Matrix

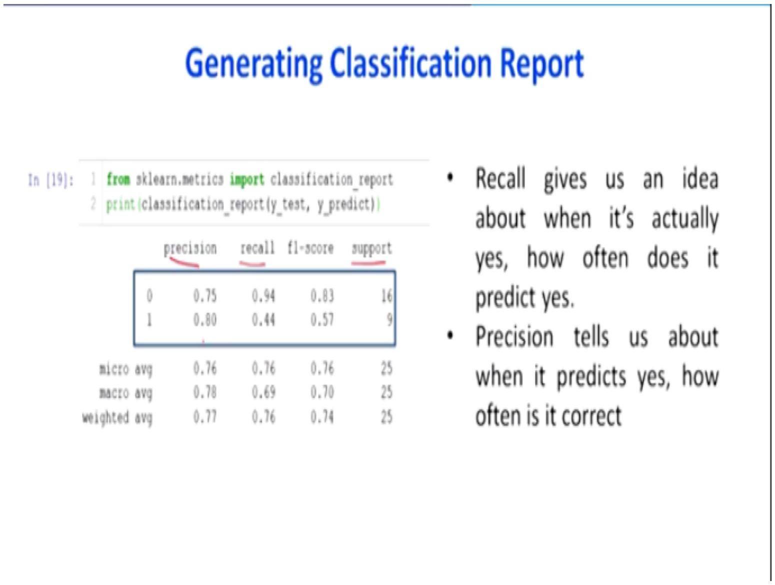
```
In [16]:  from sklearn.metrics import accuracy_score
          score =accuracy_score(y_test,y_predict)
          score

Out[16]:  0.76

In [17]:  from sklearn.metrics import confusion_matrix
          confusion_matrix(y_test, y_predict)

Out[17]:  array([[15,  1],
                 [ 5,  4]], dtype=int64)

In [18]:  tn, fp, fn, tp = confusion_matrix(y_test, y_predict).ravel()
          print("True Negatives: ",tn)
          print("False Positives: ",fp)
          print("False Negatives: ",fn)
          print("True Positives: ",tp)
```

```
True Negatives:  15
False Positives:  1
False Negatives:  5
True Positives:  4
```

Because that table is the; the previous table is the base for the confusion matrix, so here the accuracy of model is test by using this score function, so score equal to accuracy underscore

score for y test and y predicted, so the accuracy of that model is 76, to get the confusion matrix; confusion underscore matrix y underscore test y underscore predict, we are getting this confusion matrix. So, here what is meaning is the true negative is 15, true positive is 4, so the false positive is 1, false negative is 5, so this is the way to write the confusion matrix.

(Refer Slide Time: 26:31)



Generating Classification Report

```
In [19]:  from sklearn.metrics import classification_report
          print(classification_report(y_test, y_predict))
```

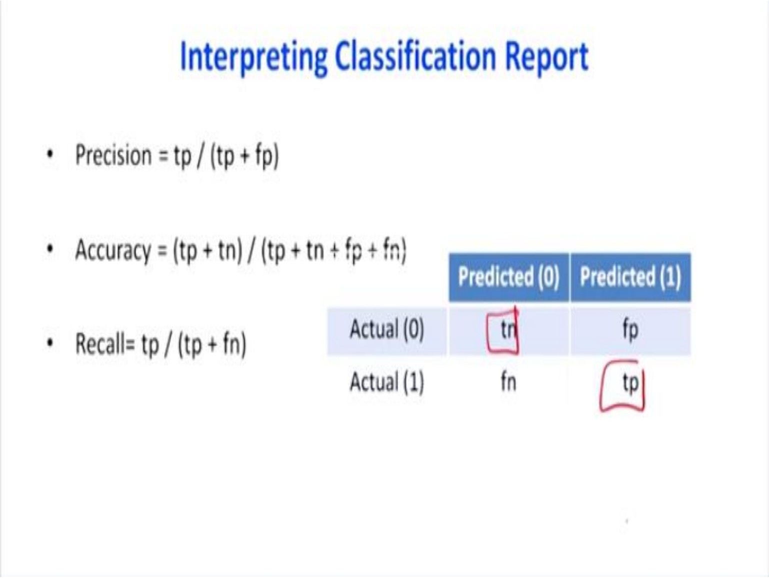
```
              precision    recall  f1-score   support

           0       0.75      0.94      0.83        16
           1       0.80      0.44      0.57         9

   micro avg       0.76      0.76      0.76        25
   macro avg       0.78      0.69      0.70        25
weighted avg       0.77      0.76      0.74        25
```

- Recall gives us an idea about when it’s actually yes, how often does it predict yes.
- Precision tells us about when it predicts yes, how often is it correct

Generating classification report; you see that when you use this functional classification underscore report, we will get this output, this is there are different columns, see one is on precision, another one is a recall, another one is f1 score and support. This recall gives us an idea about when it is actually yes, how often it predicts yes, it is like our sensitivity. Precision tells us about when it is predict yes, how often is it correct.

I have explained what is the recall, so recall gives us an idea about when it is actually yes, how often does it predict yes, it take care both specificity and sensitivity. If it is 1, we call it sensitivity, if it is 0, we call it specificity. So, in our problem the specificity is 0.94 that says that see, 94% of time we got 0 and we have predicted also it is 0, when you say sensitivity 0.44, so 44% of time, the actual is 1, we predicted also 1.

(Refer Slide Time: 27:52)



Interpreting Classification Report

- Precision = tp / (tp + fp)
- Accuracy = (tp + tn) / (tp + tn + fp + fn)
- Recall= tp / (tp + fn)

	Predicted (0)	Predicted (1)
Actual (0)	tn	fp
Actual (1)	fn	tp

So, the next one more column is the precision; the precision tells us about when it predict yes, how often is it correct, will explained this one, meaning of precision in the next slide. Now, we will interpret the classification report which was the; in the previous slide I have shown that output. The precision is true positive divided by true positive plus false positive, the accuracy is here there is one; so this cell is accurate value, this cell also accurate value.

So, sum of these 2 divided by sum of all the cells, so recall is a true positive divide by true positive plus false negative. In this lecture I have explained graphically what is ROC curve and how to choose a ROC curve, with the help of some pictures. At the end, I have taken one problem, there in that problem I explained what is a confusion matrix, how to interpret each cell in the confusion matrix. In the next class, we will continue and I will explain how to choose ROC value with the help of this same example that we will see in the next class, thank you.