Business Analytics & Data Mining Modeling Using R

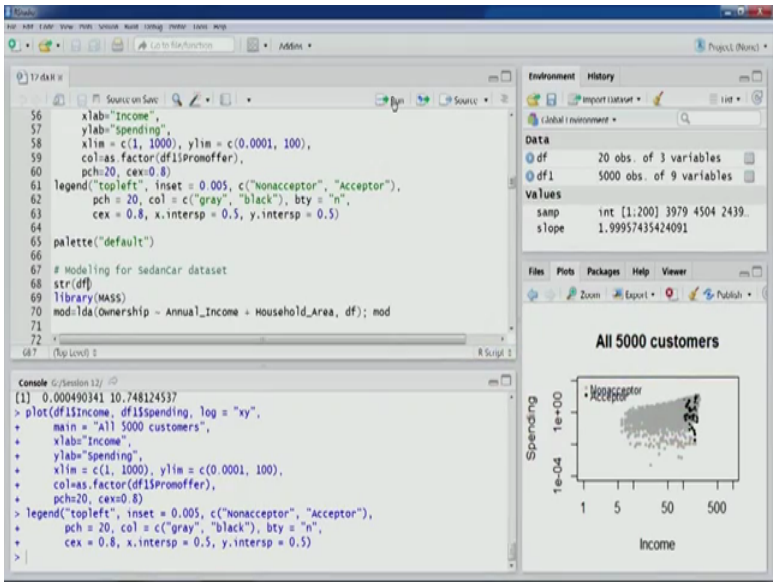
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture - 60
Discriminant Analysis-Part II

Welcome to the course Business Analytics and Data Mining Modeling Using R. So, in a previous lecture we were discussing discriminant analysis and we were a doing a few exercises in R. So, let us get back to our R studio. So, we talked about the class separation and two data sets we used to show that how class separation is important and different data sets, how it is going to impact the modeling process and results.

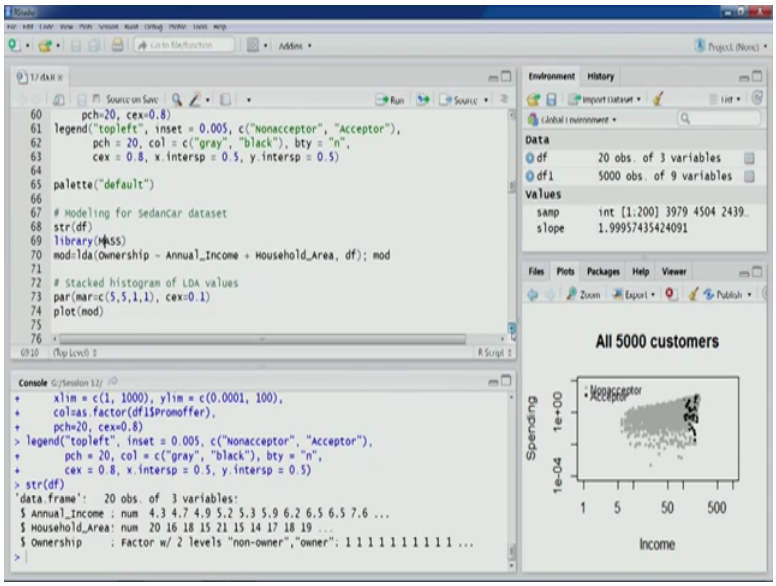
Now, what we are going to do in this particular lecture we will do our modeling. So, a we are using sedan car data set here. So, let us look at the structure of the data frame.

(Refer Slide Time: 00:58)



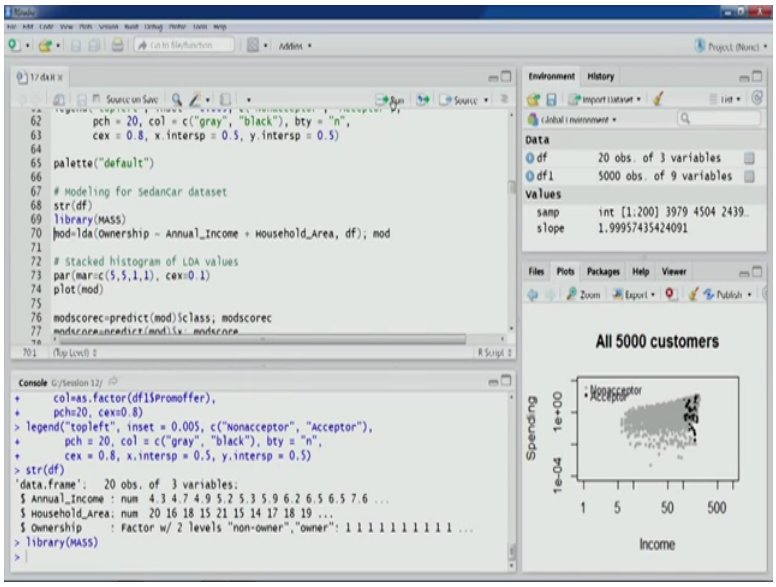
So, this is already loaded into the environment. So, you can see. So, these are the variables annual income household area and ownership.

(Refer Slide Time: 01:11)



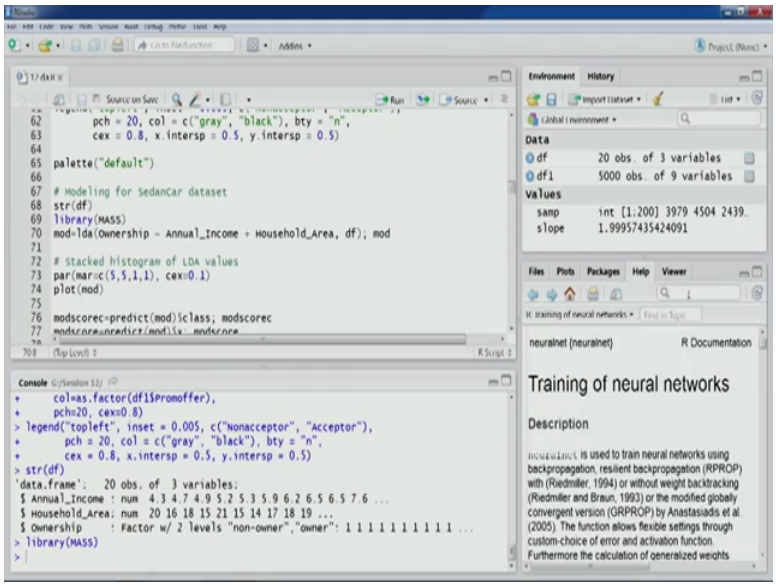
So, we can go ahead with our modeling process. So, the package that we are going to use for this particular discriminant analysis modeling is mass. So, will let us load this package library mass.

(Refer Slide Time: 01:23)



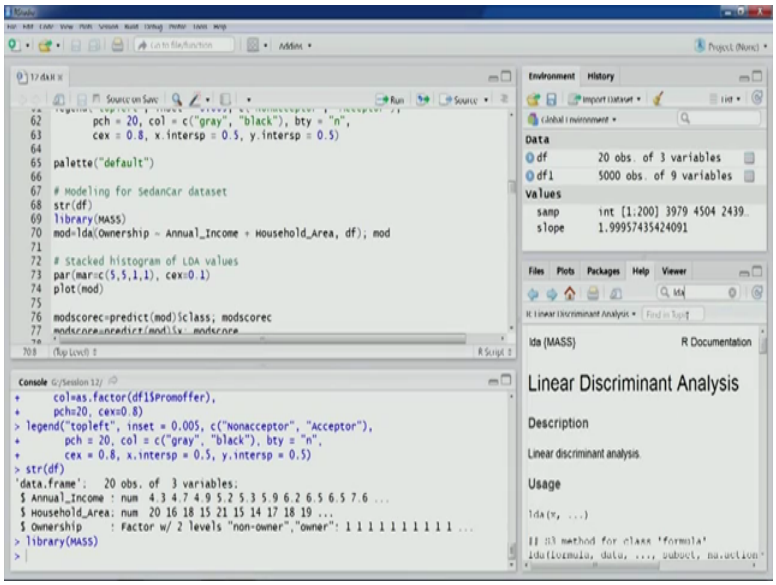
And the function that we required to build our model is called LDA. So, this is for linear discriminant analysis.

(Refer Slide Time: 01:32)



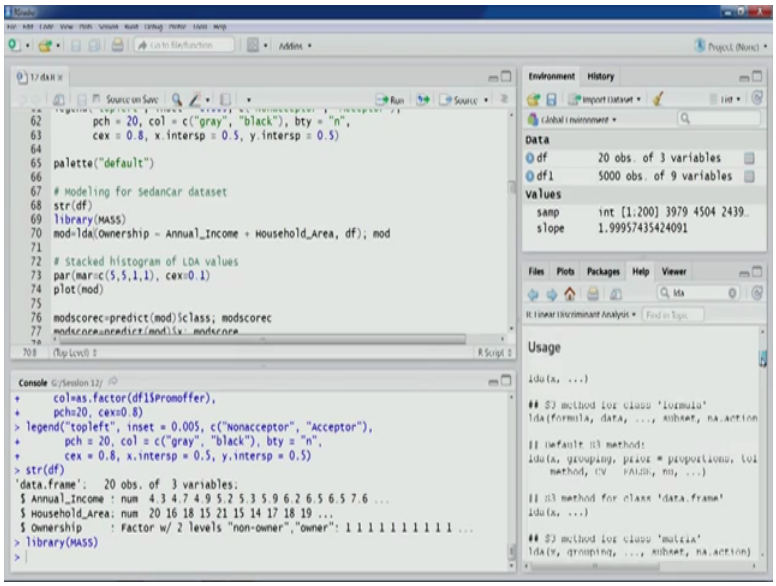
So, for more information on LDA you can go to the help section

(Refer Slide Time: 01:38)

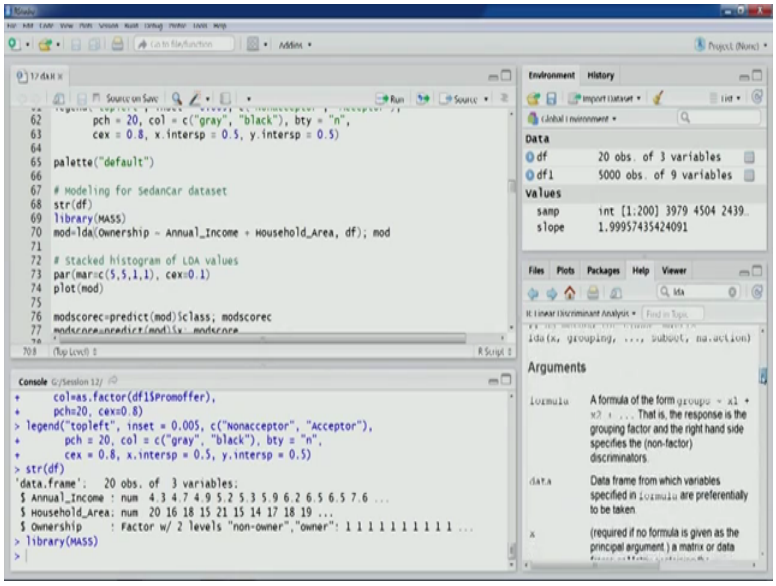


And find out few more details for example, this function is for linear discriminant analysis as you can see, and you can understand details about different arguments that are part of this function.

(Refer Slide Time: 01:44)

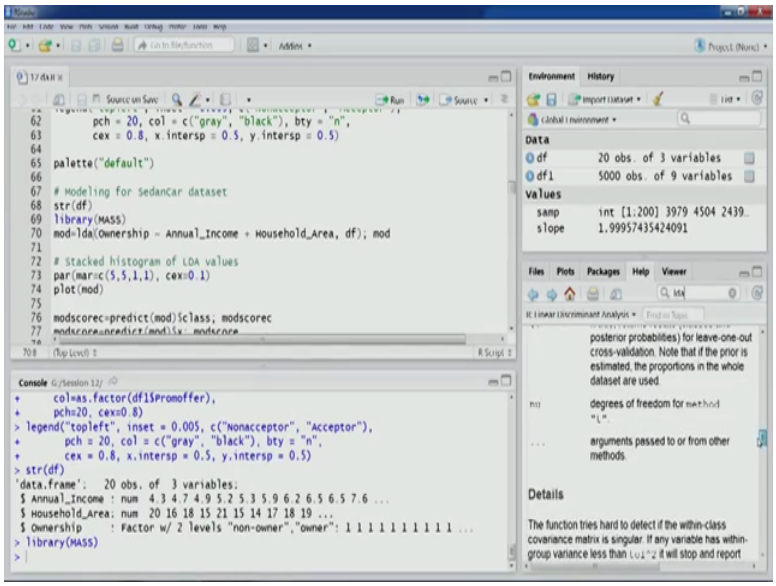


(Refer Slide Time: 01:51)



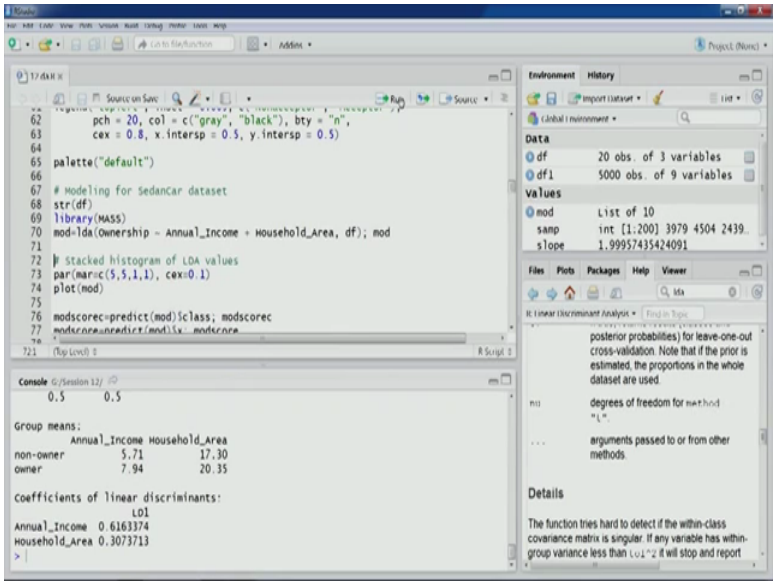
So, some of them we are going to use the important arguments.

(Refer Slide Time: 01:56)



We are going to use in our modeling exercise. So, our outcome variable here of interest here is the ownership. So, you can see ownership tilde remaining variables that are annual income and household area, and then data frame that is df and let us run this model.

(Refer Slide Time: 02:17)

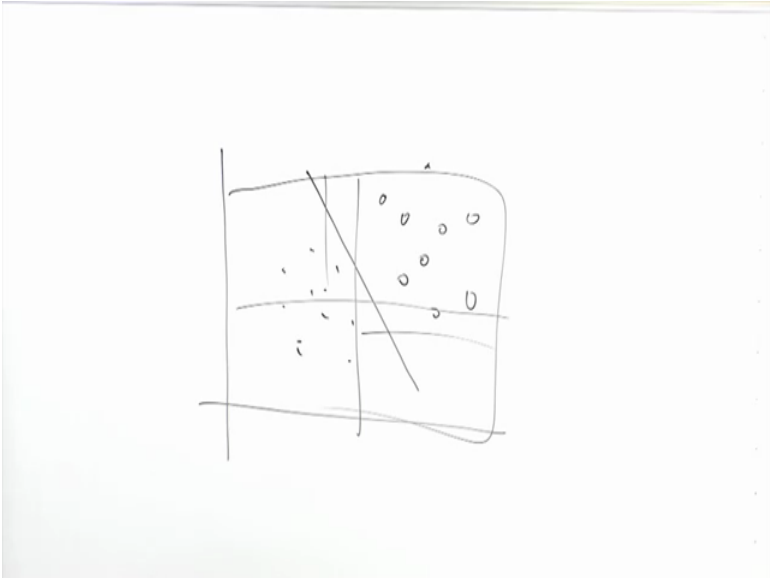


So, these are the results. So, this what call you can see the prior probabilities of groups 0.5 and 0.5 this is nothing, but the actual proportion of actual proportion of observations for each group, and then we have group means.

So, these are actually the centroids for you know. So, first row is for a centroid for non owner group and the second row is the centroid for owner group. So, the values for these two centroids, centriod you can see here. Then what we have is coefficients of linear discriminant. So, this is quite similar to what we discussed multiple linear regression. So, just like the betas that we compute in linear regression, but the idea is slightly different there that is with respect to outcome variable.

But here the coefficient of linear discriminants are with respect to the class separation, we want to achieve the class separation. So, these are the coefficient you can see 0.61 annual income and 0.3 household area. So, our LD1 that linear or function of predictors is this one this LD1 and these are the coefficients. So, this is the line that we talked about in the previous lecture as well.

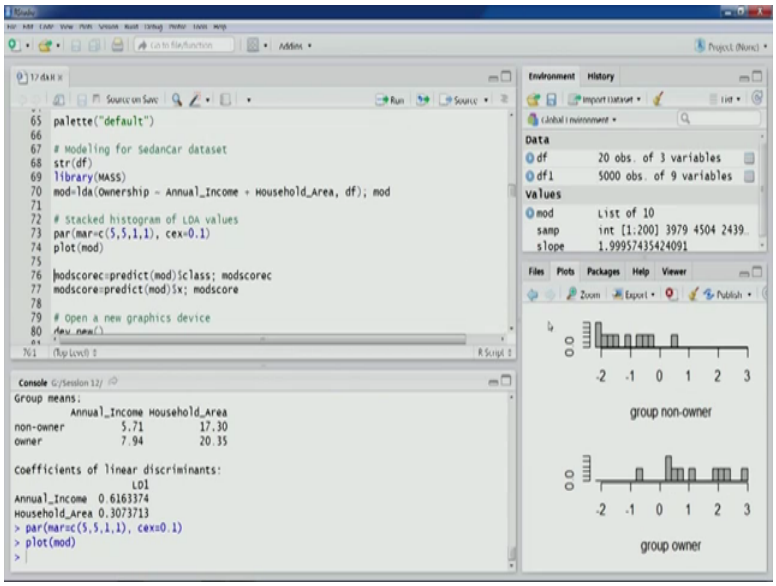
(Refer Slide Time: 03:43)



So, if we have. So, in this particular case the data set that were that we are using is also quite similar to the example that I am using here on board and this is the line that we are looking for. And the coefficients for this line we can see in this output model output, 0.61 annual income and household area 0.3.

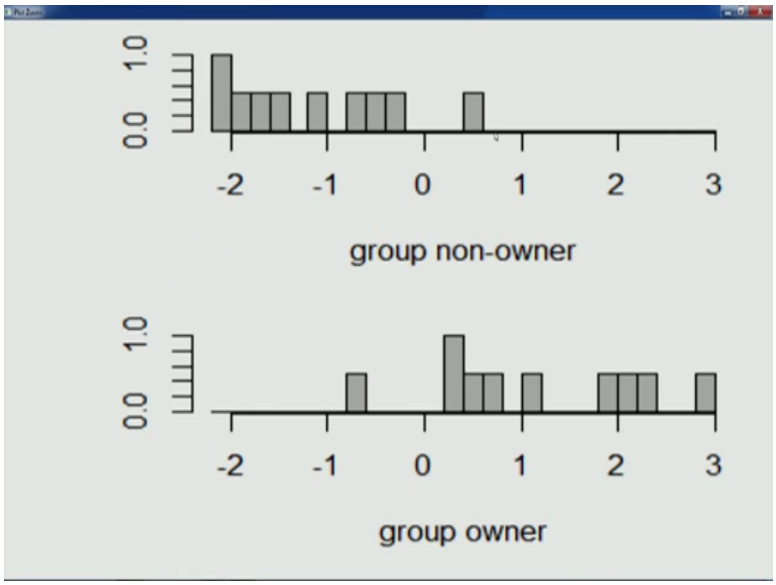
So, these coefficient will determine this line and this could be used to discriminate the observations into their respective groups. To understand the results of this model a bit further, we can use a stacked histogram of LDA values for the observation that we have. So, let us set the parameters first.

(Refer Slide Time: 04:41)



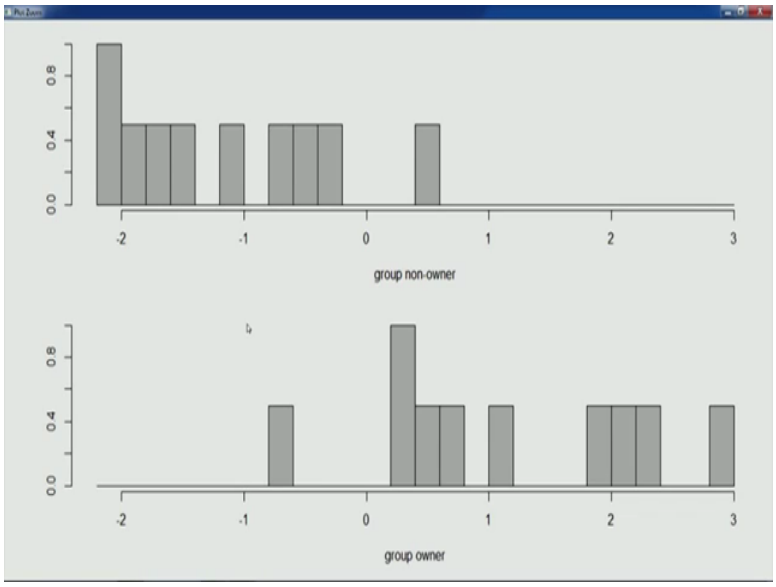
So, graphical parameters and then let us plot this and we will see this graph here.

(Refer Slide Time: 04:44)



So, in this graph as we can see that the LDA values for different groups.

(Refer Slide Time: 04:45)

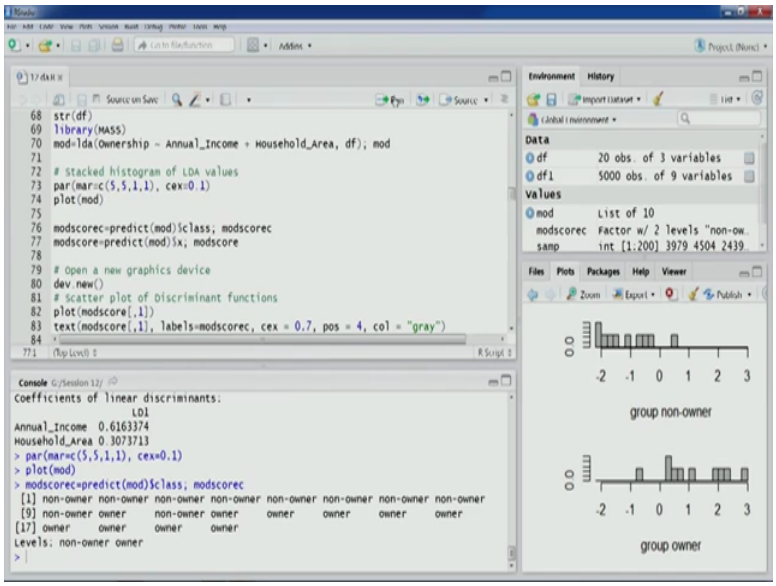


So, these are actually different observations we can see that in the data set we had just 20 observations. So, you can see here 4 5 and then 9 here and then we have 4 and about 8 9. So, almost about all the observations are covered here and we have LDA's values and we can see for the group non owner most of the values are below 0 except 1 and then we have a LDA values for the group owner most of the values are greater than 0 you can see here. So, these are so, this in this fashion our model is able to discriminate.

So, the idea that we talked about that we are looking for a line, we want to find the line that would be able to discriminate the observations into their respective groups. So, you can see that LDA's values less than 0 typically it is non-owner group and greater than 0 typically it is owner group of course, there are going to be few misclassifications.

So, this is the one understanding of the model that we have just filled. So, look few more details for example, the class that is going to be predicted using the so, this is for the training partition, the observations that have been used to train the model. So, we can use predict function and this is going to return us a one particular value that is class for each observation which is predicted class.

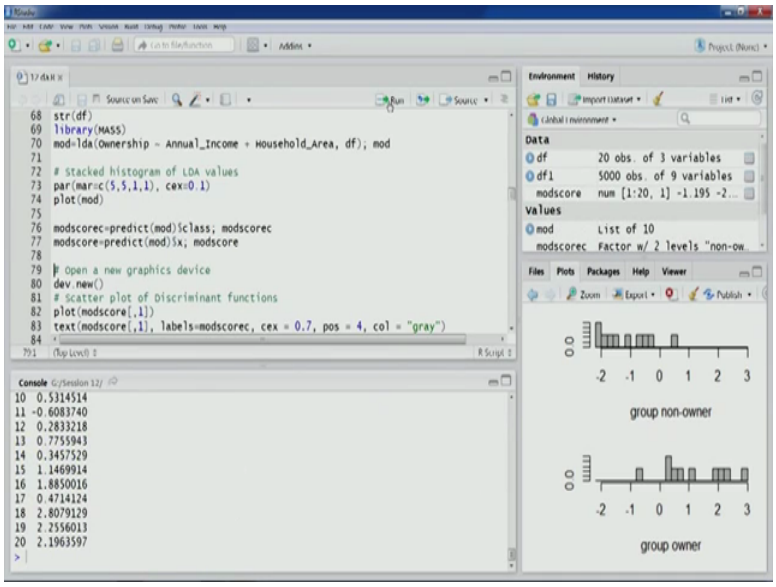
(Refer Slide Time: 06:33)



So, let us look at this. So, for each of the observations we have about 20 there and the predictor class we can see here right. And then its course also the classification is course that we talked about. So, this linear discriminant you know a function the classification linear classification function will give us these scores also.

So, predict function again can be used and this is the element one of the retained value and that can be used to displace scores.

(Refer Slide Time: 07:06)

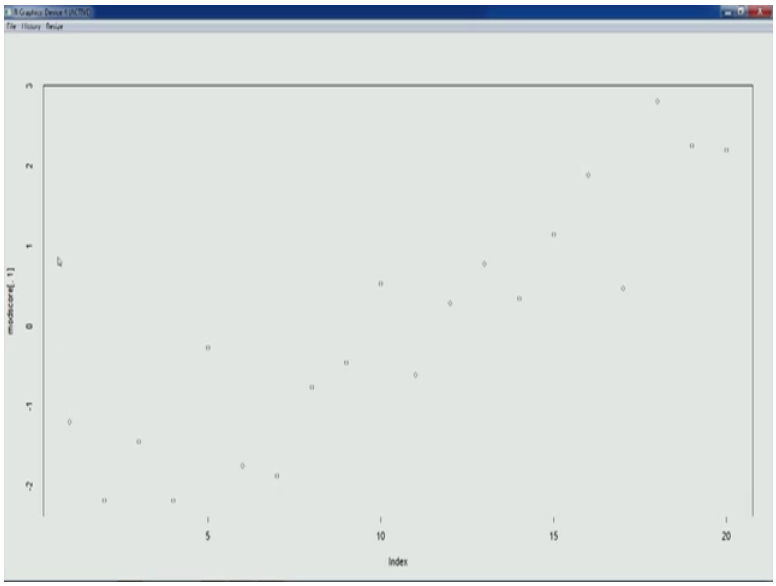


So, you can see here 20 a classification scores. So, as we saw in the plot itself that for the non-owner group the values were less than 0, the same thing you can see here, you know less than 0 values, first observation second normal observation up to you know up to a 9 observation ninth observations all values are negative. And then we can see tenth observation, it has positive value here and then we can see here then other observation they have greater than so, their values are greater than 0.

So, we can see 11th is of course, seems to be misclassification, negative then we can see positive values there 0.28, 0.77, 0.34 and in this fashion we keep on going. So, these are the values. So, we can see here, one see there seems to be one misclassification in each case. So, we can see observation 11, that is misclassified here and similarly there also this one the observation 10; that is misclassified there in the non-owner group and this 11th observation misclassified in the owner group. So, these are the two miss classifications and classification of this code we can see here.

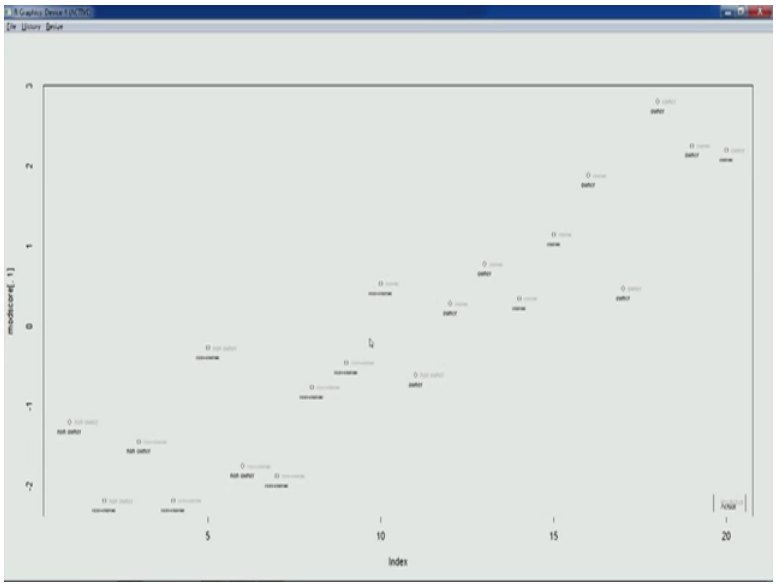
Now, if we are interested in plotting these scores so, we can plot. So, it is discriminant functions. So, in this case just one. So, we can actually plot this gate and a scatter plot for this. So, let us open a new graphics device. So, this is a function again that can be used dev dot new to open a new plot in device. So, this is how it can be done. So, this is the new device and now the plot that we are going to create this scatter plot of discriminant function, it is going to be on plotted on this device you can see here.

(Refer Slide Time: 09:11)



So, these are the classification scores for all the observations. So, you can see index. So, this index is for a row number for each of the observation, we have about 20 observations; 20 observation. So, for all 20 observations and we have the classification scores you can see, few scores are below 0 and about half of the this course are observation are scored below 0 and about half of the observation are scored as above 0 as we have seen in other plots. So, let us add few more details levels and legend.

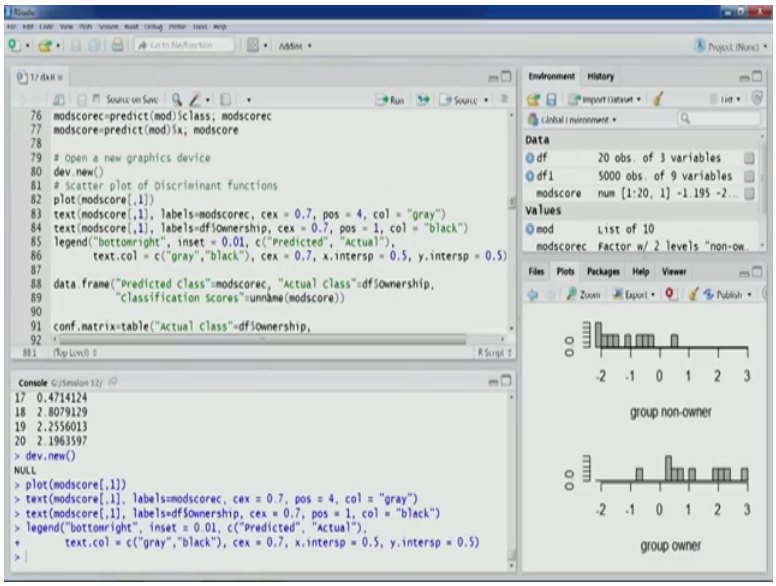
(Refer Slide Time: 09:48)



So, let us look at this observation. So, here we can see that the predicted classes is displayed using the gray colour and the actual class is displayed using the black colour. So, for each of the observation we can see here what was the actual class and what was the predicted class. Again, predicted class in gray colour and the actual class in black colour. So, we can see all these observation are correctly classified. So, we can come here and these are the observations you know, this is the observation which is actual classes owner and has been classified as non-owner here and then another one there is one more misclassification we need to find that point.

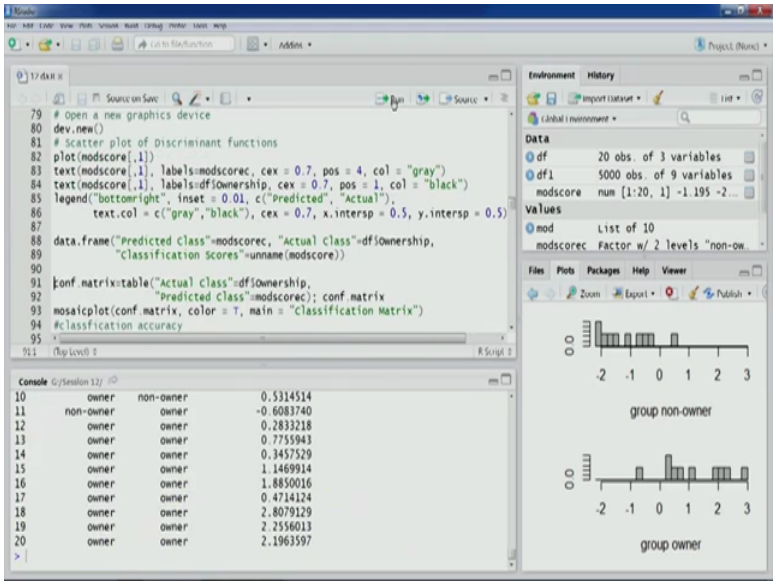
So, that is also this. So, this is the observations, we can see this first non owner and has been classified as owner. So, these are the two observations which are misclassified and so, these are quite close to our might is going to be quite close to our a discriminant line also and other observations are being correctly classified.

(Refer Slide Time: 11:00)



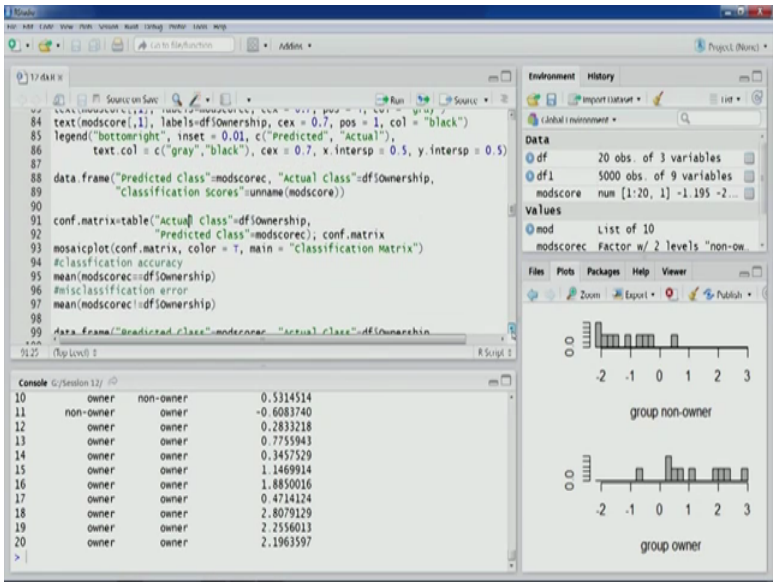
So, we can also construct a data frame of these values that we have just computed. So, let us look at some of these values in a data frame format.

(Refer Slide Time: 11:06)



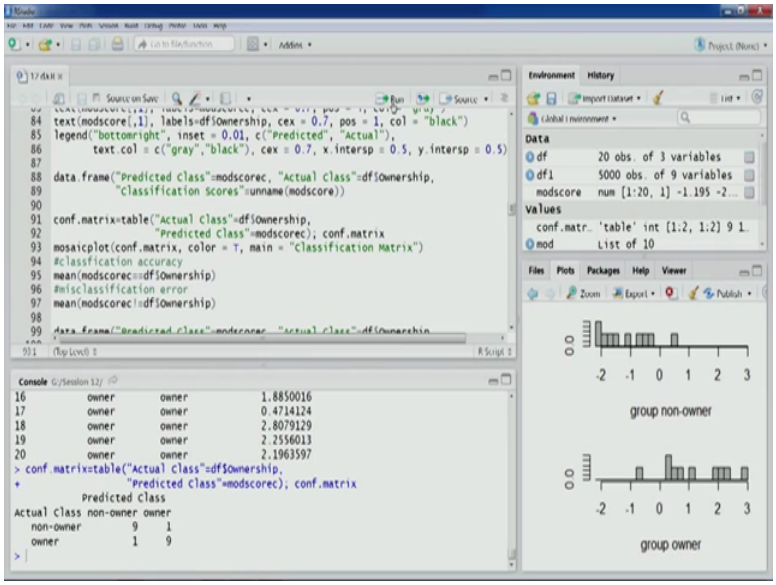
So, predicted class and the actual class and the classification scores. So, we can see a first state observations, 9 observation correctly classified. Then as I talked about, this is the first misclassification then followed by second misclassification. So, in each group we had we have one misclassification and then the remaining observation are correctly classified into its group that is owner.

(Refer Slide Time: 11:39)



Now, let us look at the performance of this model using the accuracy or misclassification error matrix. So, let us compute this matrix first.

(Refer Slide Time: 11:46)

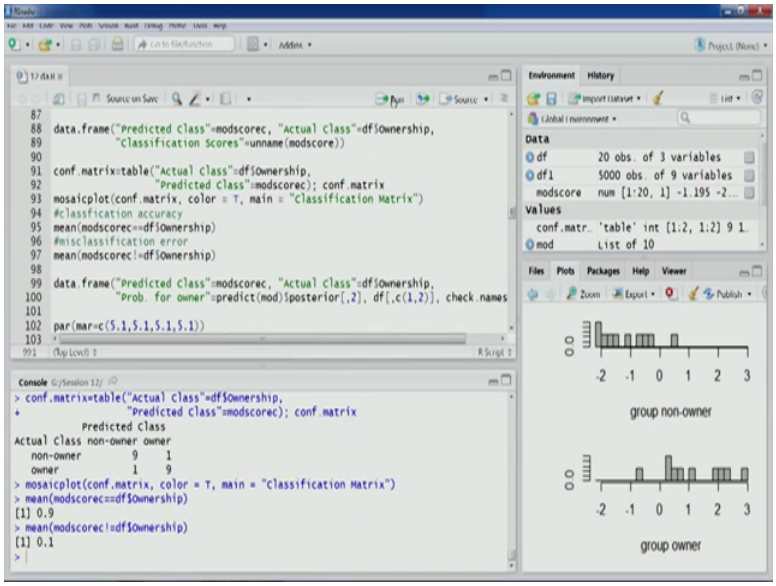


Classification matrix you can see that here 9 observations, 9 records belonging to non-owner class, non-owner class, have been correctly classified as non-owner class members and similarly for 9 owner class members have been correctly classified as owner class members and two of diagonal elements we have 1 and 1. So, these are the errors. So, let us compute the accuracy and error numbers here, before that there is

another important function that can be again used to visually expect the classification matrix, this called mosaic plot. So this can also be used to see the classification matrix.

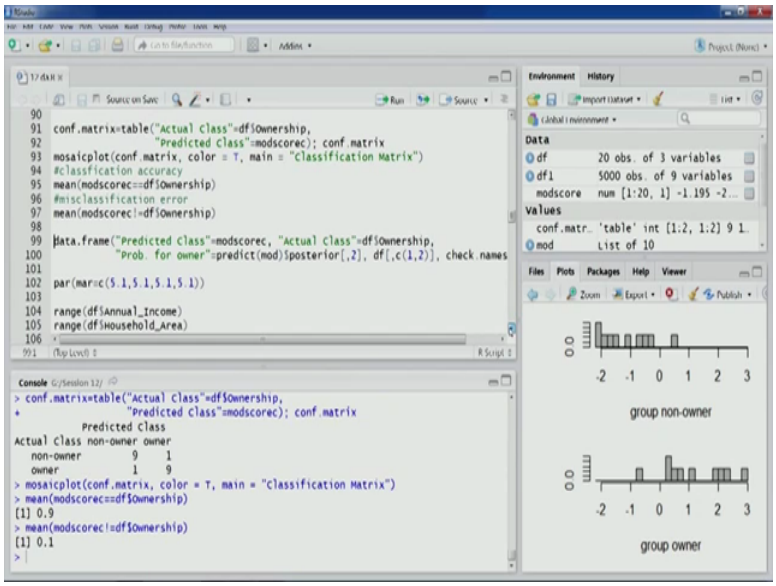
So, I think it has gone to the yes in this new device we can see. So, this is the classification matrix, then formation we can see here. So, non-owner we can see you know because the size of the rectangle, this is quite big. So, most of the observation are there correctly classified and here this size of rectangle is also quite large. So, this most of the observation have been correctly classified as owner. So, these are the observations, this be rectangles are representing a smaller rectangles are representing the incorrect classification.

(Refer Slide Time: 13:16)



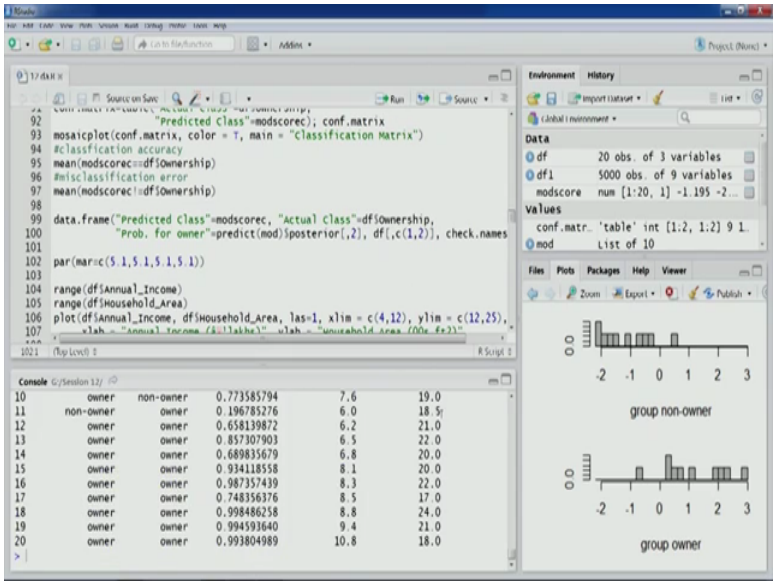
Now let us compute the accuracy and error. So, we can see 0.9 is accuracy and remaining 0.1 is the error in this case. So, model seems to be a doing a good job.

(Refer Slide Time: 13:23)



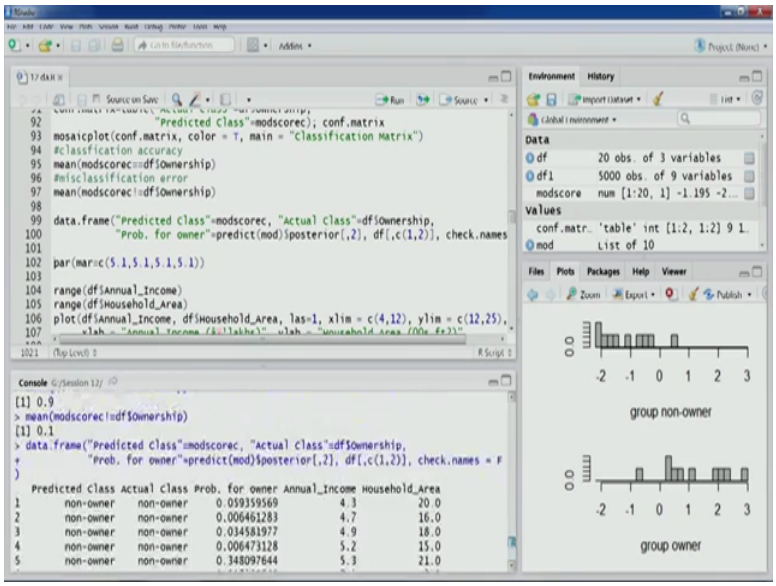
However, as we saw that in the scatter plots itself the class separation was quite clear. So, therefore, we should have expected a good performance by the model.

(Refer Slide Time: 13:37)



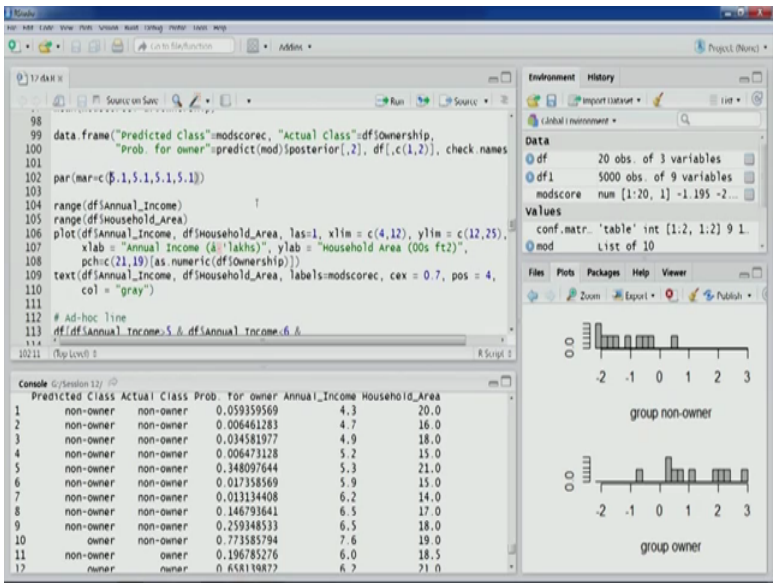
Another data frame with some key information here. So, we can see here.

(Refer Slide Time: 13:39)



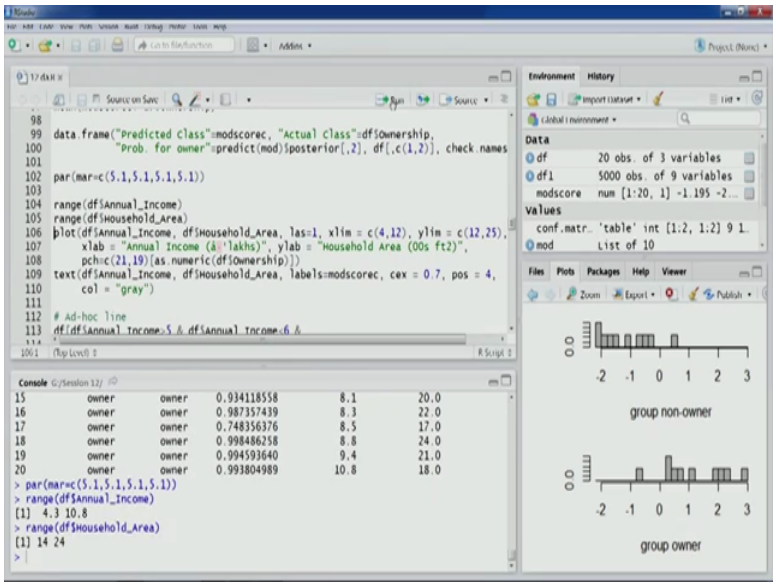
A predicted class, actual class and probability of ownership. So, that has been computed. So, you can see here probability of ownership we have this predict function and posterior element is being used to you know to actually display the probability of belonging to the owner group and the 1 minus this probability is going to be the probability belonging to the non-owner group and. So, other predictors other column names are also appended to this particular data frame. So, we have this probability values, the predicted information and predicted and actual class.

(Refer Slide Time: 14:21)

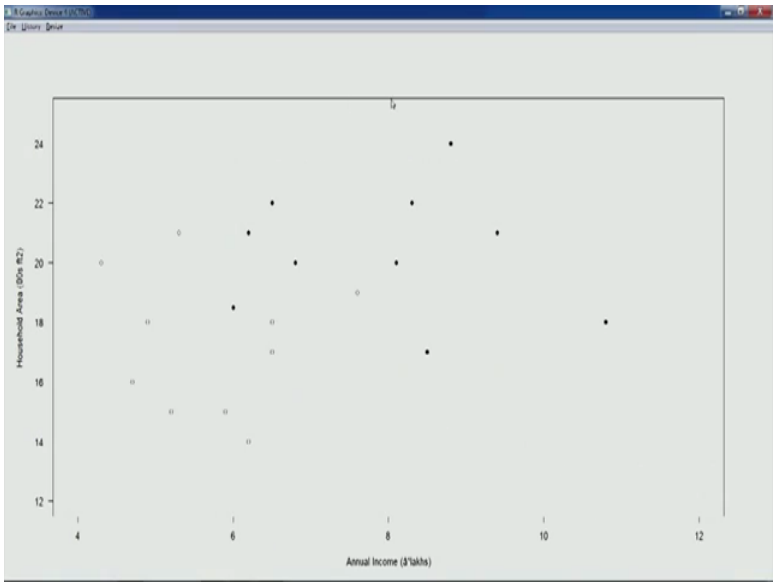


So, let us move forward. Now what will do? We will create a scatter plot where will again will create this scatter plot. So, this is again going into new device and will a use labelling.

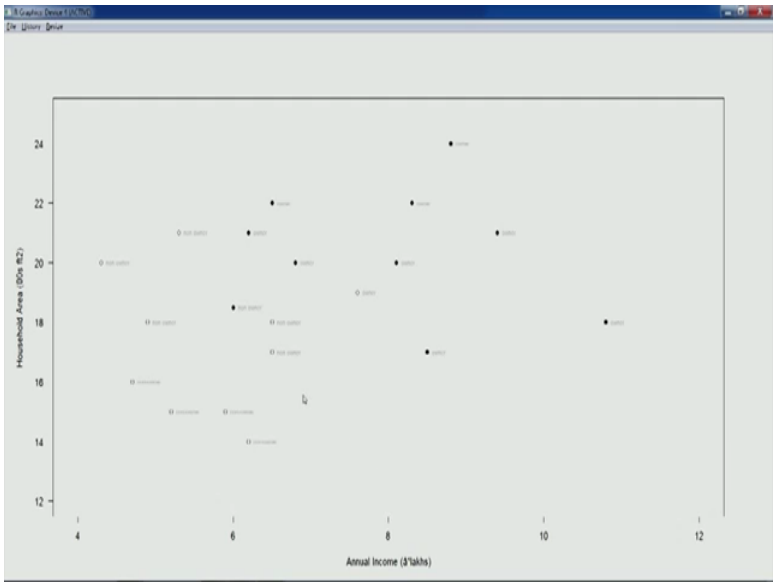
(Refer Slide Time: 14:32)



(Refer Slide Time: 14:35)



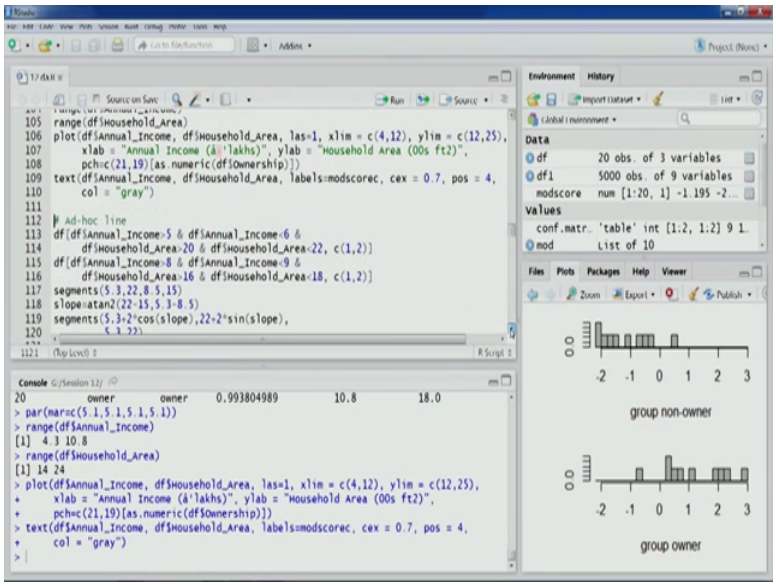
(Refer Slide Time: 14:42)



For so, you can see this is the scatter plot and we are using labelling here for different observations, and how they have been classified. See it text there so, that is telling us the predicted class.

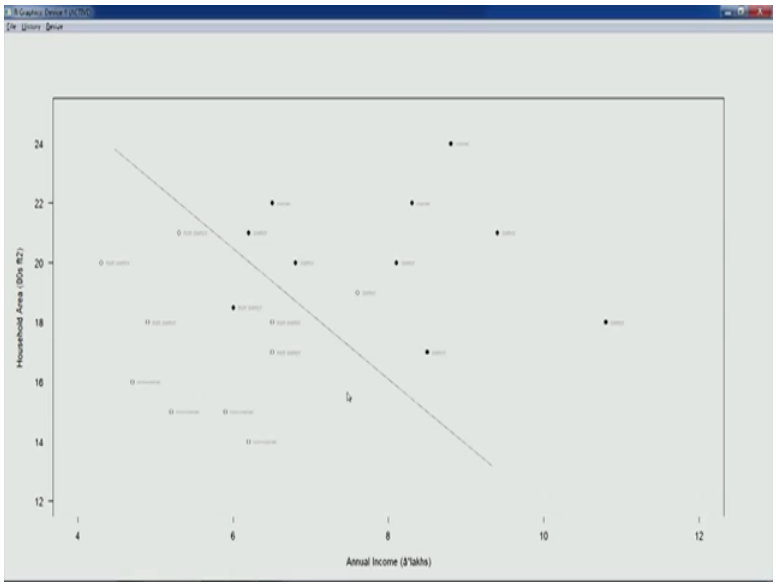
So, here again we can see that the, this particular observation. So, it was belong to the owner class, but it has been predicted as non-owner. So, this is one misclassification. And we have one observation here and this was a so, this is correctly classified and this is another observation. So, this was supposed to be non-owner, but it has been predicted as owner. So, again in this scatter plot itself, we can see which observations because this was a small data set.

(Refer Slide Time: 15:32)



We can easily spot which observation have been correctly classified and which have been incorrectly classified. Now, what we are going to do is, we again re-plot our ad hoc line that we did before in the you know previous lecture, and we will plot the discriminant line that, we have just computed using our model LDA function. So, let us re-plot. So, these computations we are already we already understood.

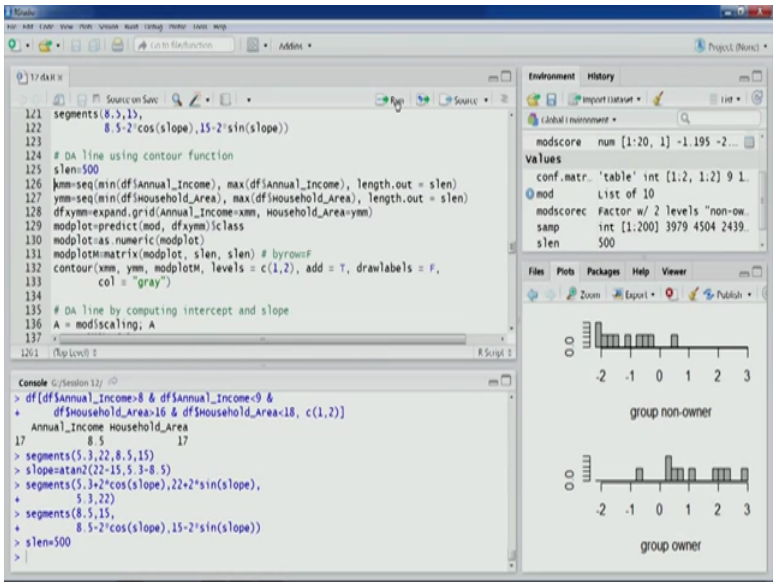
(Refer Slide Time: 16:03)



So, this is the ad hoc line that we had plotted in previous lecture.

So, now let us look at the line that we have got from the model. So, there are two ways, first one we can use this contour function. So, we would be required to do some of these computations. So, because.

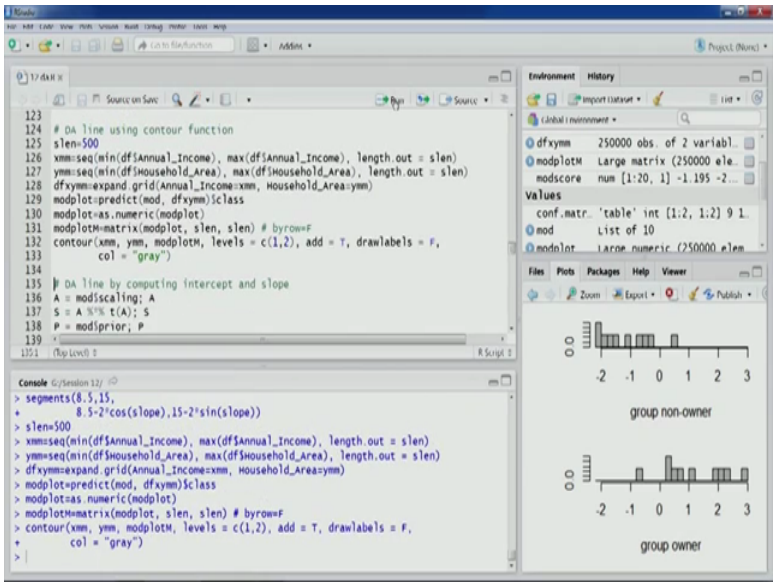
(Refer Slide Time: 16:26)



So we will have to expand our observation. So, you can see here I am using I am expanding some of these observation using expand dot grid function; more details on expand dot grid you can find out using the help section.

So, now using the predict function and the model object, I am going to score these this expanded grid this expanded data frame and now this is going to be used to create a matrix, and then create the line our discriminant line.

(Refer Slide Time: 16:54).



So, we can see here this is our discriminant line shown in gray colour. So, this is the product of the. So, this is from the model. So, this is the line that has been computed, which is actually giving us the classification.

Now, the same discriminant line can also be plotted using some matrix algebra computations. So, for example, coefficient values and other things that we are going to compute slope and so, finally, we want to compute slope and intercept for the line. So, some of these computations are related to this. So, will just go through these computations. So, this matrix of algebra is bit you know complicated here.

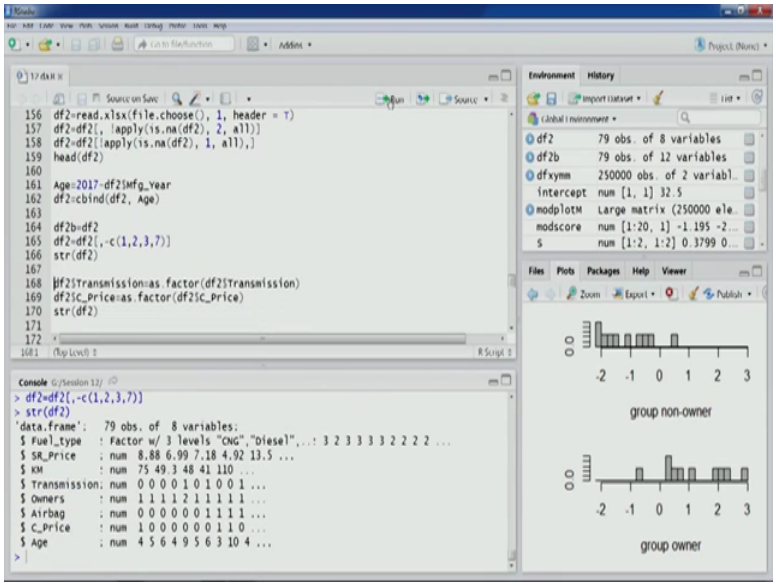
So, we would not go into detail of this, but this is another suppose DA line that is by computing intercept and slope and you can see that matrix algebra is involved, you can see that coefficient values and then we take multiplied with this transpose matrix take prior values and centroid values and then we compute our slope and intercept values. And then the intercept in the plot that we had created, let us clip is a function that can be restricted, that can restrict the plotting of a particular you know a line or any other graphic to a to a limited region in the in the graphic device.

So, let us look at. So, now, you can see that gray line because this AB line you can see colour is blue. So, this gray line you can see here this has been you know you know coloured with blue; the same gray line has converted to blue because the DA line is going to be the same it is just different plotting mechanism that we had used.

So, you can see here from the legend itself also. So, earlier the gray line that was there that was the DA line, now it is blue, this is the DA line, this is our black line, this is our ad hoc line, that we had drawn on our own by looking at the you know observations and other details are there.

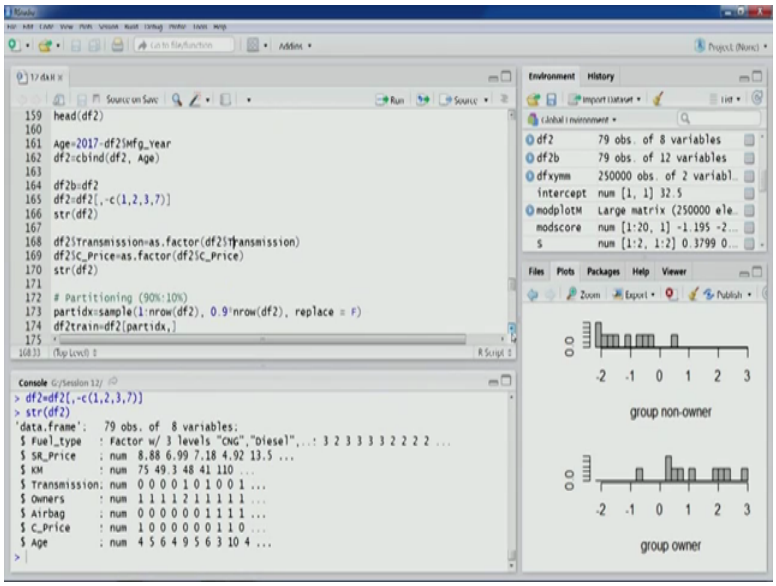
So, let us move to another exercise, now this time what we are going to do is we are going to use a this particular data set used cars data sets, let us import this data set. So, let us move na columns na rows. So, these are the observations now let us compute each variable, let us take backup and again you can see that this we are going to build a classification model here.

(Refer Slide Time: 19:41)



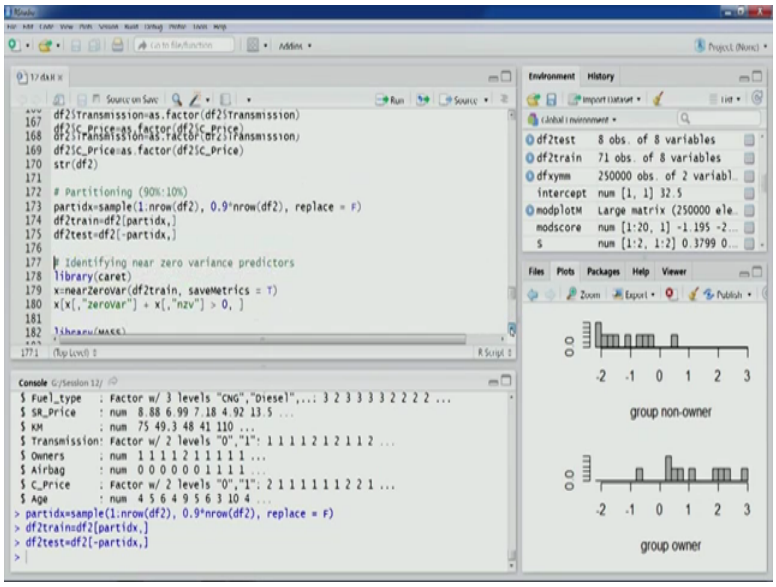
So, we are eliminating the price variable and will just keep the C underscore price, as you can see here this is our factor variable categorical variable and let us convert to other variables into factor transmission and C price.

(Refer Slide Time: 19:50)



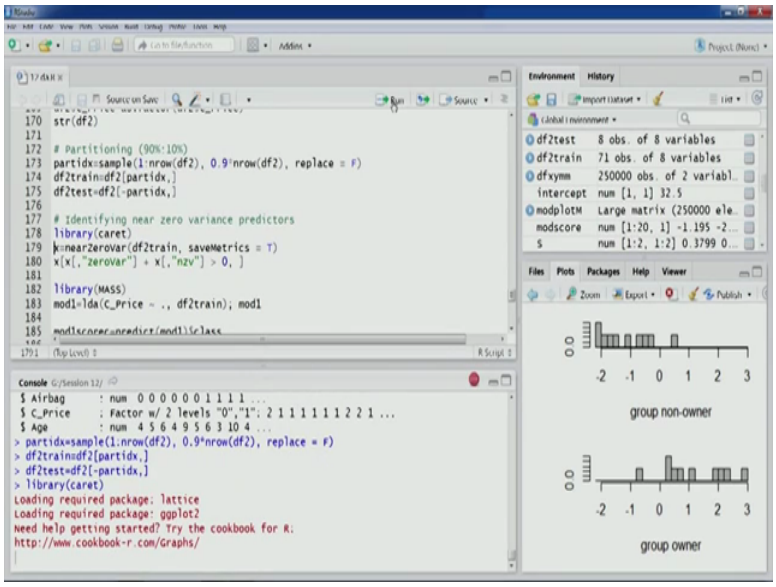
So, this is our final structure. So, we can see here, now we can actually partition. So, let us partition so, 90 percent of observation to training and testing. So, you can see in the earlier exercise all the observations were part of the you know part of the model and now this time we have slightly bigger samples. So, we are creating different partition and then we would be judging the performance on the validation or test partition.

(Refer Slide Time: 20:25)



So, this is another important function that that is near 0 where so, this is to identify near 0 variance predictors. So, ideally we would not like to include these near 0 variance predictors in our LDA modeling.

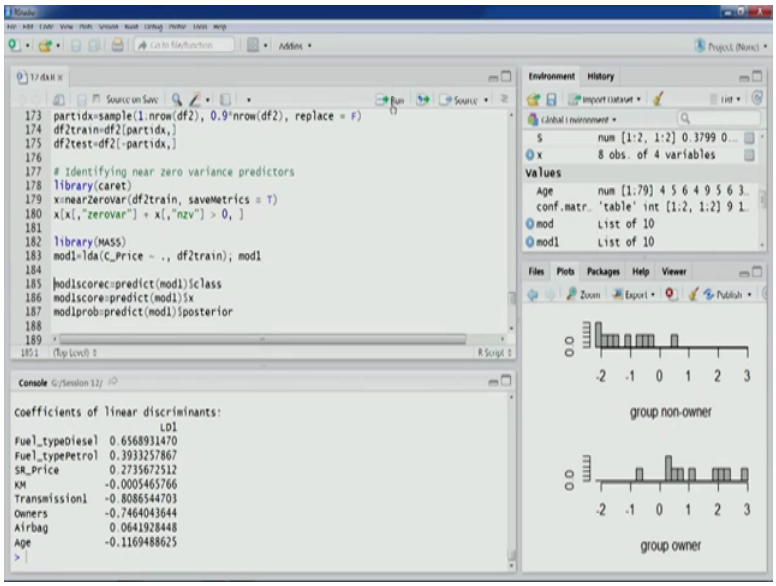
(Refer Slide Time: 20:47)



So, library this caret library is there and which has this function near 0 variance to which spot these variables predictors.

So, we can see in our case we do not have any you know near 0 variance predictors. So, we are safe we can go ahead and do our modeling. So, this was the library that we has used for LDA exercise and LDA is the function on this case you can say C price being request against all other predictors and this is our training partition.

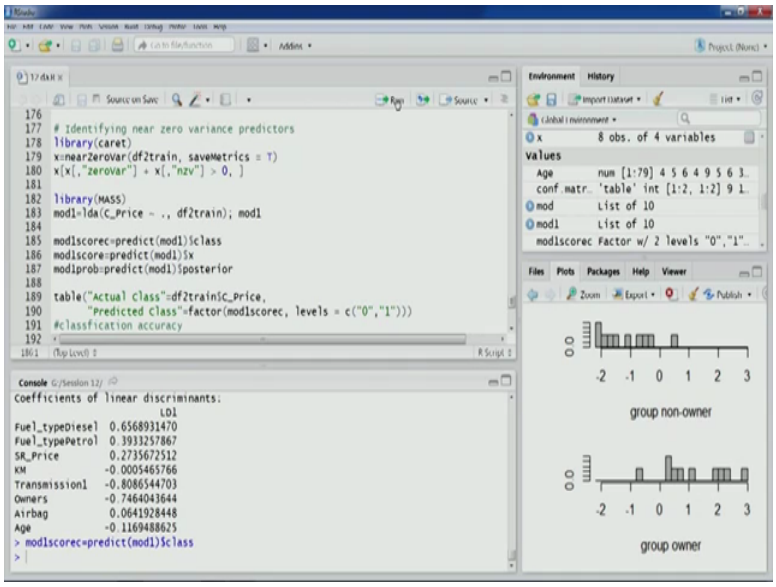
(Refer Slide Time: 21:19)



So, let us build the model. Now you can see we since we have just two groups here again. So, we just need one you know linear discriminant function here and we can see the combinations, this is combination linear function of all the predictors.

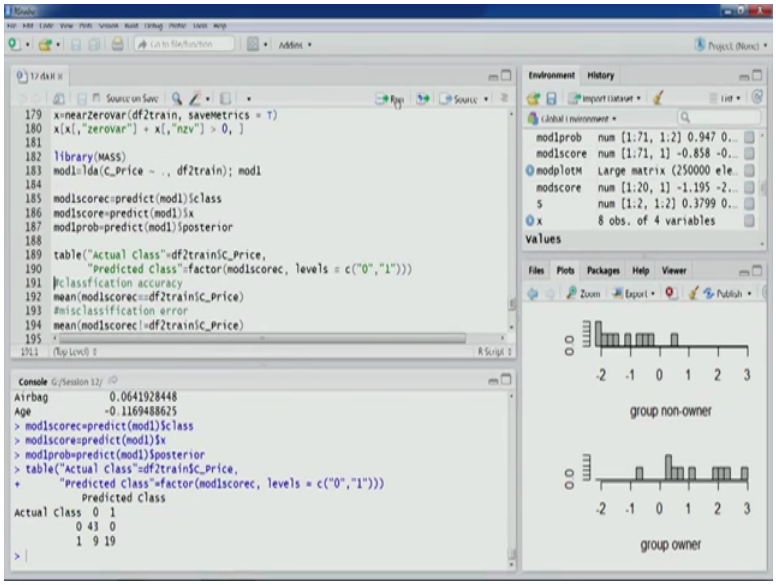
So, you can see diesel, petrol, SR price, KM transmission, owners, airbag, age. So, all these variables we can see here and we can see the our linear discriminant function, now we can go ahead and it compute some of these scores.

(Refer Slide Time: 21:55)



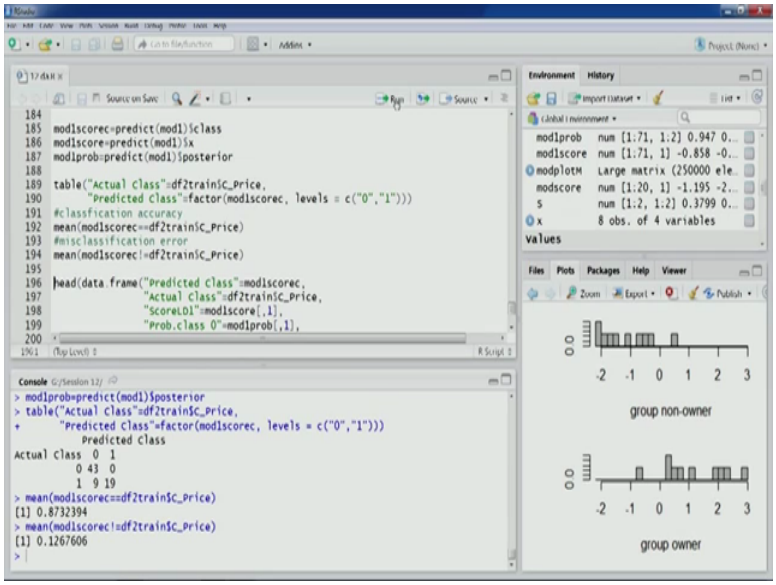
So, classification then the scores classification scores and then the estimated probabilities values and let us look at the classification matrix.

(Refer Slide Time: 22:04)



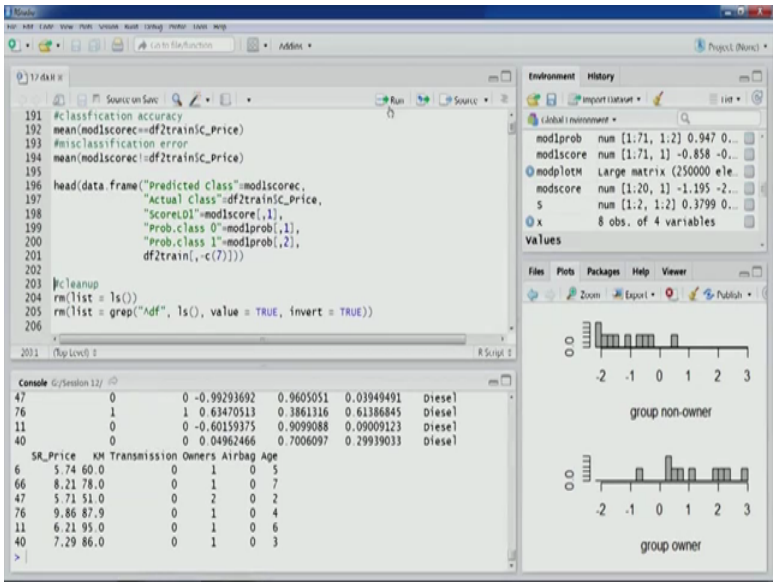
Now this is the classification matrix, this is on training partition. So, we can see the model is doing a good job here.

(Refer Slide Time: 22:15)



Let us look at the accuracy and error numbers, 87 percent and 12 percent. Now let us look at this particular data frame with some information on some key variables.

(Refer Slide Time: 22:26)

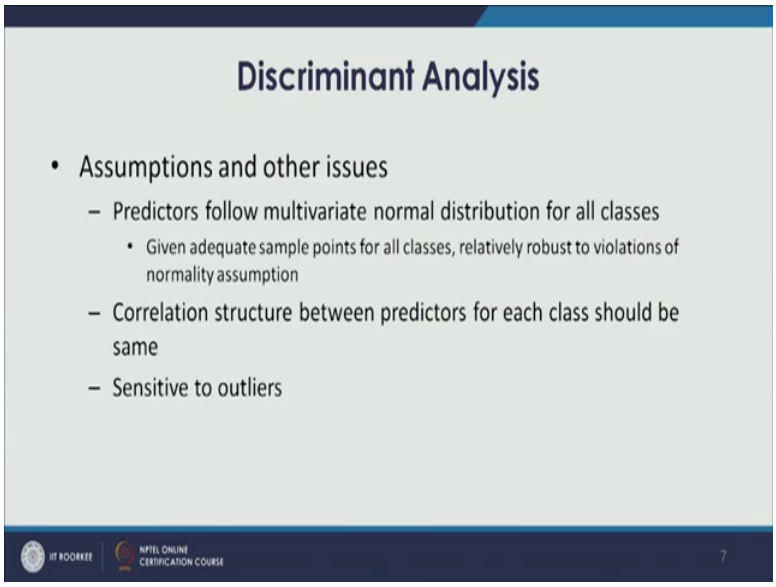


Let us look at this. So, we have predicted class, actual class, we have a score that is from coming from the our discriminant function. So, this is our classification scores, then we have probability of belonging to class 0 and probability of belonging to class 1 and after we have all the predictor. So, this particular data frame has all the important information, the output of the model and the input as well. So, this in this fashion so, as you can see that, we can we can if the class separation and other things are quite good then you would see that the 1 linear discriminant function does a quite good job. So, we can compare our scenario, this linear discriminant scenarios to the classification and regression trees.

So, there we used to identify the rectangular regions, which would be you know able to separate some of these observations in this fashion. So, in this case you would see instead of looking for a set of horizontal and vertical lines, creating rectangular region as we typically doing classification and regression trees. In this case we are looking for a diagonal line or a line, which is able to discriminate between the observation and you know classify them into their respective groups. So, each technique has its own strengths right you can see, instead of looking creating so many rectangular regions, as we do in CART. Just one discriminant line was sufficient for this particular data set to get a good model.

So, let us go back to our discussion on using slides. So, few important things that we would like to cover here. So, there are some assumptions and other issues that are concerning discriminant analysis.

(Refer Slide Time: 24:23)



For example, predictors, first assumption is that predictors follow are supposed to follow multivariate normal distribution for all classes. So, the predictors that we have been that we have to used for our to construct our discriminant linear line that discriminant function.

So, for each of these those classes they should follow a normal distribution. However, if there are good enough sample points, then the results are quite robust to violation of this particular assumption, about the adequate sample size for you know each of those classes have to be there. So, this is first assumption and then the second one is that correlation structure between predictors for each class should be same. So, that structure is important for us to be able to build a valid discriminant analysis model. So, this is the second one. So, we can always look at the correlation matrix and find out whether the correlation structure is similar for each of those classes.

Now, this technique is also sensitive to outliers, as you can understand from the discussion as well, that if there are a few outliers our linear discriminant line can change significantly and that can impact the results. So, therefore, it is important to identify those extreme points and eliminate them if possible. Few more comments on

discriminant analysis so, as we talked about that application and performance aspects are similar to multiple linear regression.

So, just like in linear regression with respect to outcome variable, we compute our you know a coefficients. So, as you can see in the second point, in discriminant analysis coefficients of linear discriminant or optimized with respect to class separation. So, we want to achieve class separation and the coefficient of linear discriminant that line that are optimized in that fashion. When we talk about linear regression the coefficient are optimized with respect to outcome variable because we want to predict that outcome variable, so coefficient or output price with respect to that. So, that is one difference otherwise there are so, many similarities, only the approach is different the optimization process. So, in both cases the coefficients or weights are optimized and the estimation technique is same. So, we use least square in both the cases and discriminant analysis as well as multiple linear regression, these squares is used, the same estimation technique is used here in discriminant analysis.

So, with this we have completed our discussion on discriminant analysis and with this we have also discussed we have covered most of the supervised popular supervised learning techniques under this course. So, we started with so, first to be had introductory lectures, then we started you know understood the data mining process, we looked at some of the exploratory techniques visual visualization techniques, we also went through the matrix that are used to assess the performance of classification model and prediction model, then we started our discussion on formal techniques that are used for modelling.

We started with multiple linear regression and then KNN, naïve Bayes, we covered neural network, we also covered classification and regression trees. So many other techniques that we have covered logistic regression and discriminant analysis. So, all these techniques that I just talked about, they come under the umbrella of supervised learning algorithms. So, there is always an outcome variable and with respect to that outcome variable, we go about building our models either for prediction tasks or classification tasks.

So, in this course in this course and this part of the this particular course, we have been able to understand some of the basics of data mining modeling and analytics in general, and we covered statistical techniques, mathematical techniques, data mining machine

learning algorithms and we have understood how these techniques can be used in modeling process. And most of as I said most of these techniques are typically used for classification and prediction that is supervised learning techniques.

So, I hope that you have learned a lot you know going through the lectures of this course, and you would be able to use some of the information some of the learning's that you have used here in your future in your career.

Thank you.