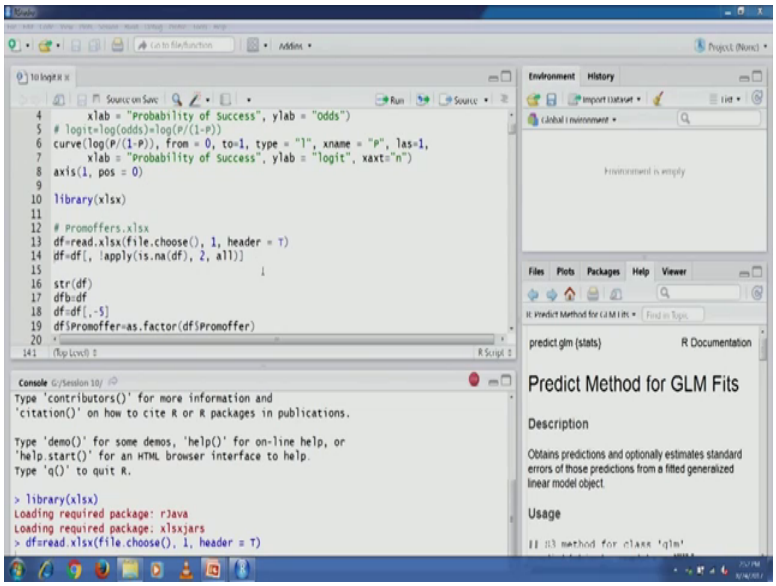
Business Analytics & Data Mining Modeling Using R

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

Lecture - 49
Logistic Regression - Part IV

Welcome to the course business analytics and data mining modeling using R. So, in previous a few lectures, we have been discussing logistic regression and in a previous lecture, specifically, we talked about how we can actually interpret a Logit model, Odds model and also probability based model, we also understood the differences in terms of interpretation . So, let us in this particular lecture let us start with our exercise in R that we have been doing. So, we have been using this promotional offers data set. So, we would like to complete this particular exercise.

(Refer Slide Time: 01:00)

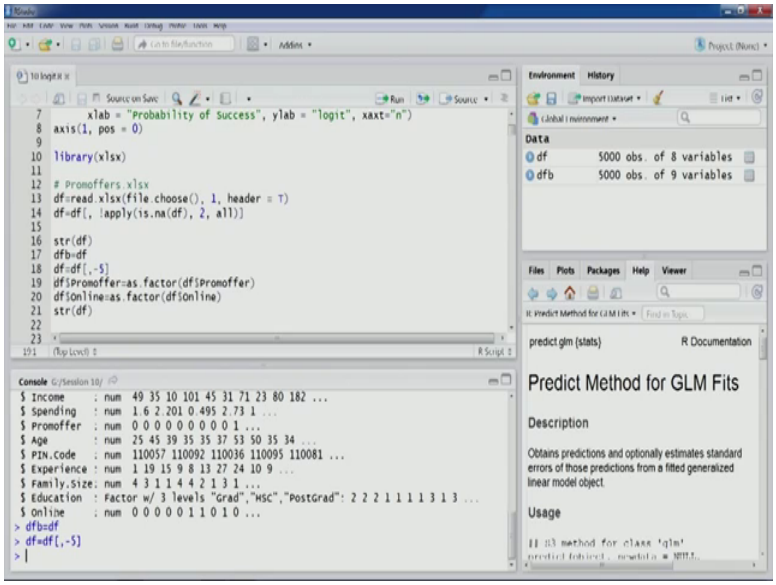


So, let us load this protocol library xlsx. So, promotional offers data set that we are already familiar with 5000 observations. So, let us a load it into R environment.

So, a in a previous lecture, we have been able to build the model and we also understood the results and interpreted the results that we got in our promotional offers model. Now we will check the performance of this particular model on test partition and also will for the training partition as well we will look at some of the charts like cumulative lift curve and also design chart for this particular data set. So, as you can see now observations

have been loaded into environment section you can see this 5000 observation; let us remove NA columns let us look at the structure once again.

(Refer Slide Time: 02:03)

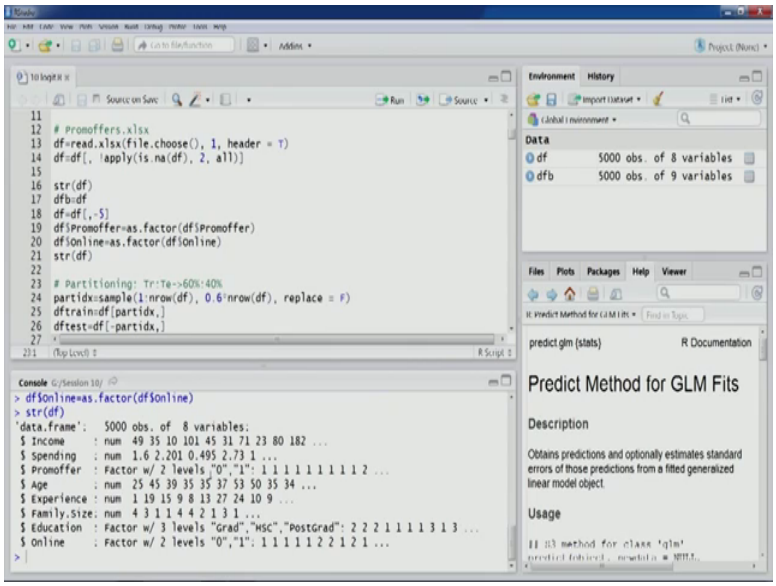


So, all the familiar variables, right.

So, let us take a backup of this particular data frame and we are as we talked about in previous lecture as well, we are not interested in this particular variable pin code and many categories, right. So, we would like we would not like to consider this particular variable in this model. So, let us get it off get rid of this particular column. Now we are left with a promote to categorical variable promotional offers and online activities online activities whether a customer whether a particular individual is active online or not and the promotional offer is our outcome variable of interest whether the customer accepts the offer on or not.

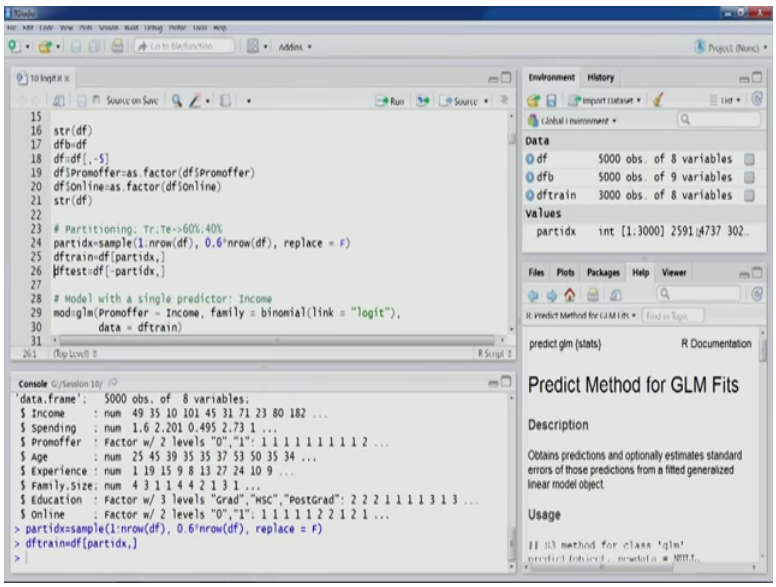
So, let us convert them a to factor variable categorical variable now these are the variables that you would like to take forward for our modeling exercise income expanding promotional offer and then age experience family size education and then online.

(Refer Slide Time: 03:00)



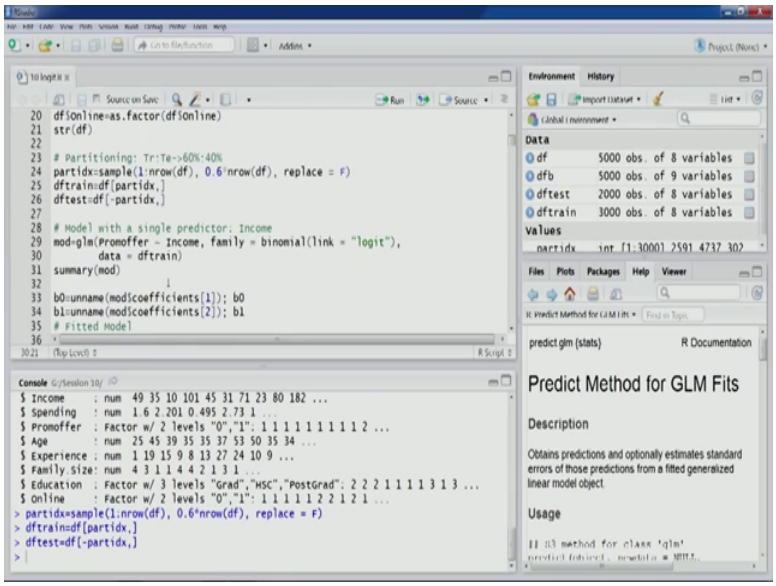
So, we followed a 60 percent, 40 percent partitioning in previous lecture as well. So, let us do the partitioning.

(Refer Slide Time: 03:20)



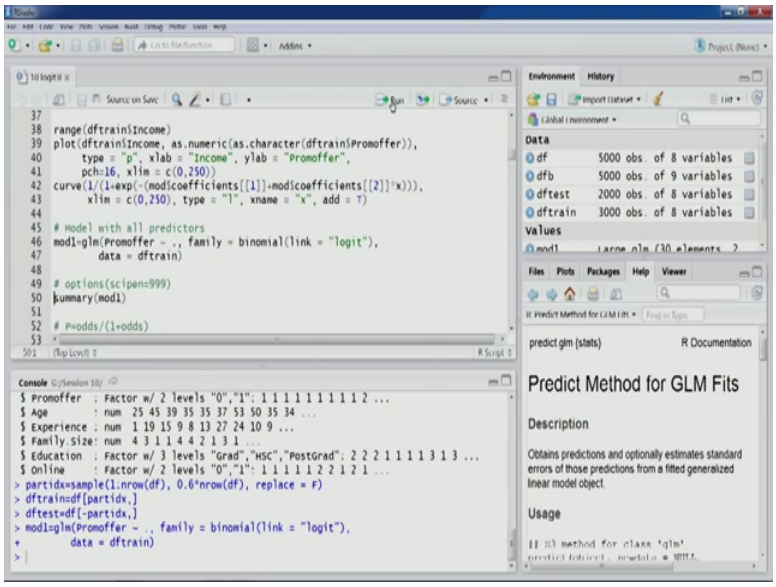
So, 60 percent of the observation will go into the training partition as you can see in the environment section 3000 observations for df train 8 variables and for test partition the remaining observations that is 2000 observations on 8 variable. So, that is also there.

(Refer Slide Time: 03:43)



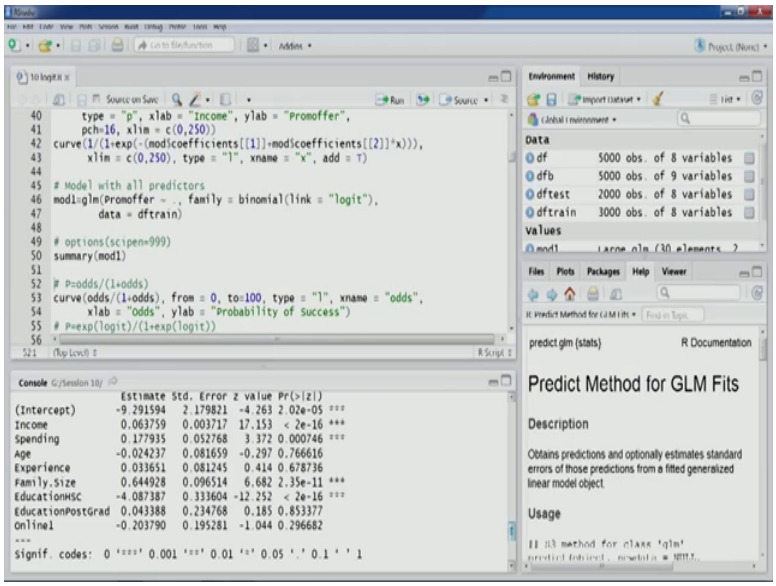
Now, as we talked about in previous lecture the GLM is the function that can be used. So, this program order we have already build model with single beta we already discussed this one.

(Refer Slide Time: 03:53)



So, let us move to the model with all predictors. So, as you can see in this particular model, we are we have this formula promotional offer tilde dot so; that means, we are going to build model against all predictors using all predictors right other parameters remain same. So, let us run this run this model, let us look at the summary.

(Refer Slide Time: 04:23)



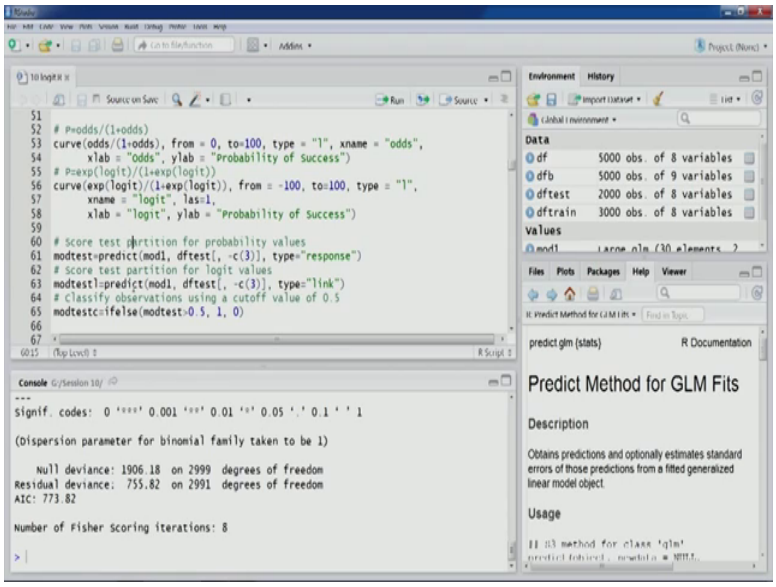
So, this particular model also the results of this model also we have discussed, however there is a one slight change in the results ah. So, in the previous run; that we had done in the last lecture, as we can see is spending this particular variable right. So, this was this particular variable this was significant at 90 percent confidence interval level in the p in the v in the run that we did in previous lecture; however, you can see that as the sample as changed ah.

Now, this is also significant at 99.9 percent significance level right. So, the results that we have today, in today's run; we can see that income is spending and the family size education HSC; these are the significant variables and 3 of them were a significant at 99.9 percent level in previous run as well and in this particular lectures, run a spending also comes out to be significant. So, this is slight; this can happen when we run a particular model multiple times. So, a larger sample size can a guarantee us a more stable results more robust results right because it the model results also depends on the observations because training partition, we randomly draw observation from the full data set and then use them for training partition.

So, the observations the every time we run the observation a that are used for model a are going to change and therefore, a slight differ slight differences in terms of a model can be seen to repeat a to repeat the model a using the same observation as we have talked about in some of the initial lectures of this course set dot seed function can be used set

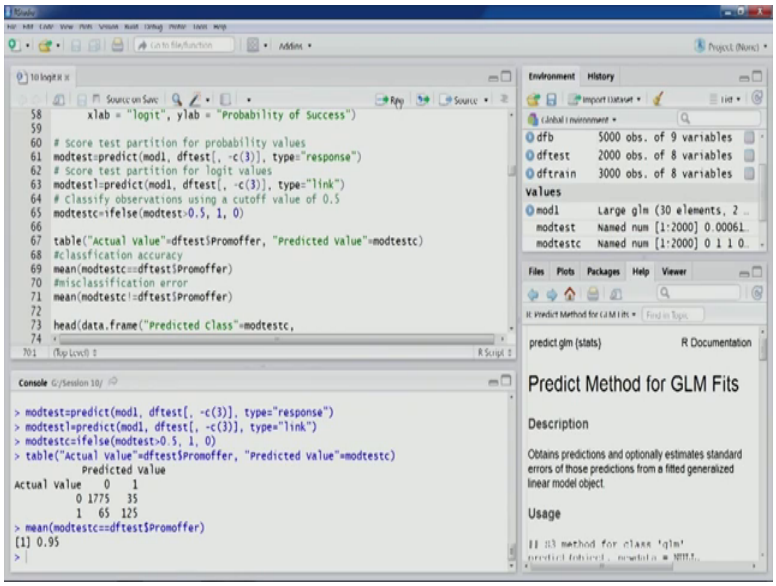
dot seed function will actually allow us to use the same partitioning same observations pertaining partition for the modeling as well. So, we have already discussed the results of this particular model. Now let us a move forward. Now let us check the performance .

(Refer Slide Time: 06:39)



So, test partition ah. So, we will like to score test partition for probabilities values. So, this is how we can do it predict function. So, this particular aspect also we talked about we need to this third argument type we need to specify as response to have probabilities values and this particular argument will has to be specified as link to have the Logit values and then we will have to manually classify the observations based on the probabilities values right. So, let us score the probabilities value Logit values.

(Refer Slide Time: 07:22)



And then we can score in this fashion the observations. So, cutoff value is 0.5. So, we have just two classes. So, this is a two class case. So, 0.5 cutoff value of 0.5 will be equivalent to most probable class method and in this particular case. So, let us use it.

So, now let us look at our classification matrix. So, with this code we would be able to generate the same. So, you can see in the classification matrix out of the 2000 observations that we have in the trend test partition as you can see in the environment section as well. So, out of 2000 observations that we have 70, 175 observations have been correctly classified as class 0 members and hundred and twenty five observations have been correctly classified as class one members the of diagonal elements that is 65 and 35. So, these are the observations which have been incorrectly classified either into class 0 or class 1. So, we can go ahead and compute our classification accuracy. So, this comes out to be 95 percent in this particular run.

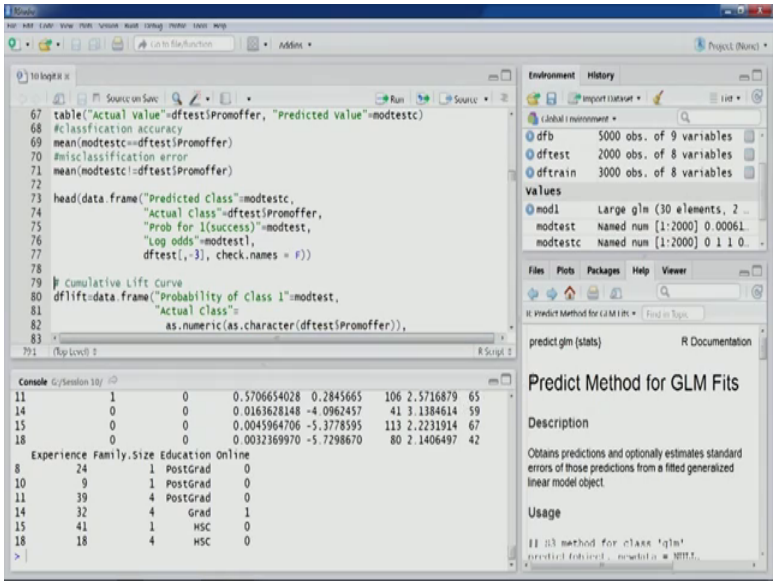
If you remember in the previous run that we did in a in the last lecture there also we got the similar number. So, that was also near about 95 point something in last lecture. So, you can see the model is a on in terms of performance numbers in terms of matrix the model is quite this stable and robust right in previous run we also got similar perform. So, the remaining is the error that is 5 percent.

Now we can compute the important variables for this particular modeling exercise where you have a predicted class actual class predicted class is stored in mod test c and then we

can create a data frame of all these important key variables here actual class is stored in promotional offer we can have probability value mod test. So, using this we can also have a look at the table this particular data frame and the table and have a look how our model has performed log Odds also we can have in this fashion mod test l that we have already computed.

And this is our then we can also have the test partitions those variables here in this particular data frame. So, let us look at first 6 observations of this particular data frame.

(Refer Slide Time: 10:00)



So, you can see predicted class and the actual class. So, because our accuracy is 95 percent for this particular model; however, in first 6 observational itself you would see that one error for this particular observation the actual class was 0, but it has been predicted as one, if we look at the a probability value for the same we can see that 0.57 is the probabilities value. So, is the probability value? So, therefore, it has been classified as class 1 ah; however, actual class is 0.

If we look at the other numbers for example, the first row here you can see the probabilities value is quite low. So, therefore, it has been correctly classified as class 0 we look at the row number 2 the probability value is .98 quite close to 1. So, it has been correctly classified as class 1 and this one is the error right probability value is a more than 0.5 therefore, it has been classified as one even though the actual class was 0; if we look at 3 remaining a 3 remaining rows also, we can see that the all for all these 3 rows

the probability value is much less than 0.5, it is quite close to 0. So, therefore, all the all these 3 rows have been correctly classified as class 0.

So, a log Odds value a you can also see. So, you can see the values which are close to 0. So, as we had seen the plot of you know probability a probability versus log Odds logic values. So, from there, we also I can understand that the log Odds value is Logit values on the negative side so; that means, it will have the corresponding probabilities value quite close to 0. So, the same thing you can we can see in all the rows where the Logit values are negative similarly a positive Logit a values as we saw in previous lecture, in the plot that positive logic values.

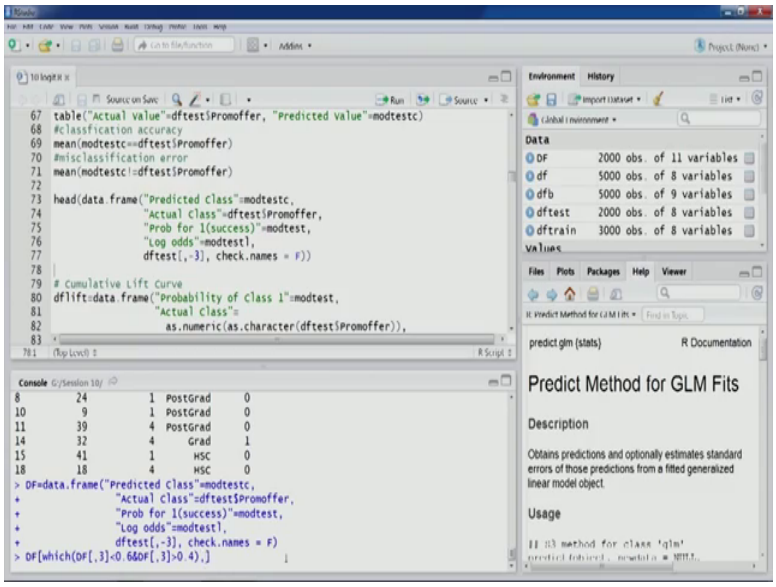
They will typically mean a higher probability value a probability value close to one the same thing is reflected in row number 2 positive Logit value and higher probability corresponding value if we look at this particular value. So, we saw that that around the you know when the Logit value is around 0 mark, then we see sudden you know change in a probability value. So, all the variation in the probabilities values come around the when the Logit value is near about 0 mark.

So, you can see 0.28 when the Logit value is near about 0 mark 0.28; you can see that the probability value is also near about 0.5, right, you can see 0.57 in this particular case and these are the cases and these are the cases the cases where Logit value is close to 0; that means, the probability value will be close to a 0.5 mark you know on either direction. So, those are the cases which will which will be difficult to classify for a model in this case, as we can see also row number 3 the model was not able to classify correctly the observations.

Then a the predictors variables have also been added to this particular table. So, that can also be analyzed accordingly income spending aged experienced family size education. So, that can also be analyzed. So, if we look at the most interesting row that is the row number 3 here you can see the income levels the spending and the age. So, on the higher side our experience and family size and education. So, we can look at different a values specific values for a particular observation and we can understand the results further another thing that is possible here is that a we can have a look at the we can have a look at the values which were which are you know which have been incorrectly you know which have been incorrectly classified by the model.

So, if you are interested in those value. So, we can previous command was this one.

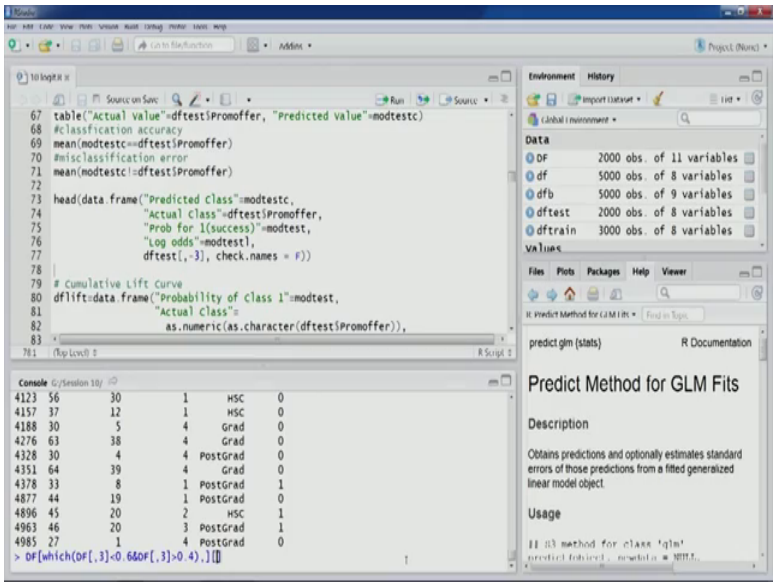
(Refer Slide Time: 14:23)



So, a in this particular data frame itself we can look for the values which have been incorrectly classified right. So, first we will have to store this particular variable in a data frame. So, let us say df. So, we; so, we store this particular variable in this a data frame df and now within this df, if we are interested in finding the rows where the predicted class was not equal to predicted class was not equal to the actual class, right or rather more interesting rows would be where the probability value a the probability value that is a the third third that is the third column right probability value is close to 0.5 right.

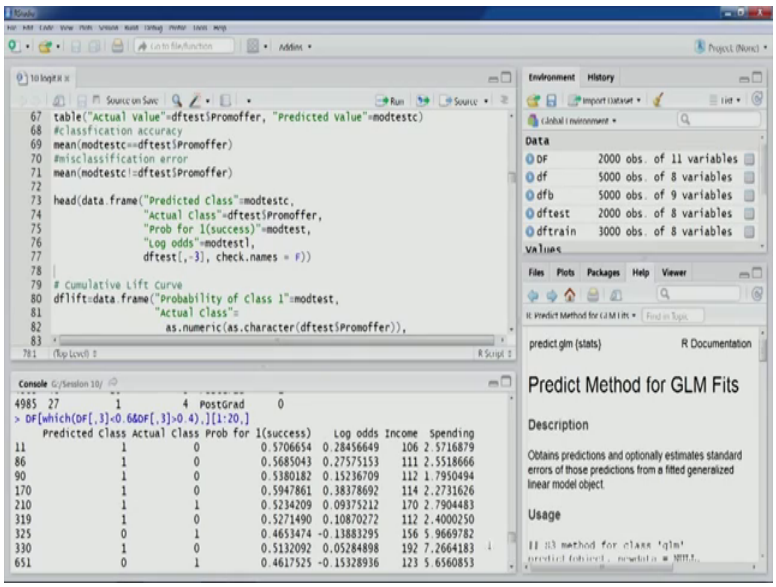
So, that would be more interesting the, those would be more interesting observation. So, let us compute the same. So, third row and we would like it to be let us say less than 0.6 and the same observations we would like it to be greater than let us say 0.4; so, all the observations which all the rows which follow this.

(Refer Slide Time: 16:00)



So, you we can see here and the results. So, that there could be too many observations in this case. So, there can be too many observations. So, let us take a first few observations let us take a let us twenty observations here again. So, in this fashion we can do it. So, let us scroll.

(Refer Slide Time: 16:30)

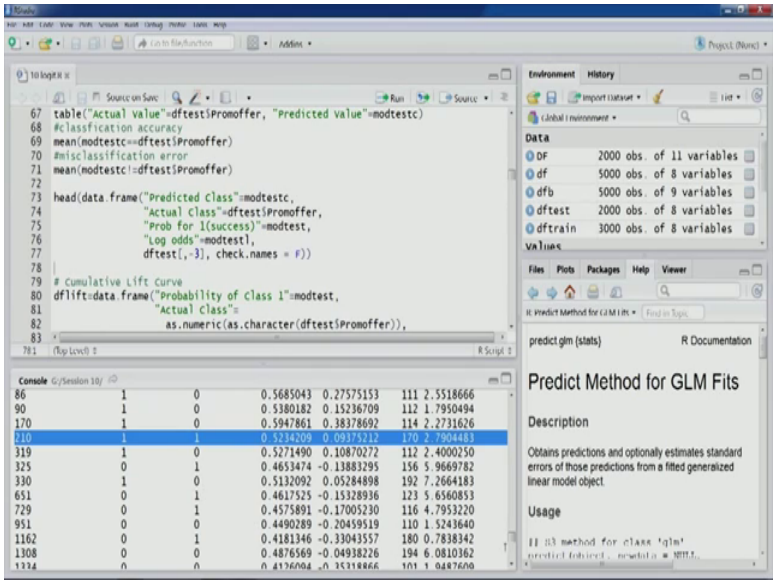


So, now, we can see so, these are the observations for which we probability values range from 0.4 to 0.5 as you can see from our criteria as well, right. So, probability is value range from 0.4 to 0.6. So, that was the range where Logit values a you know are close to

0 and we see a change in you know sudden change, you know a in a probabilities values near about this range.

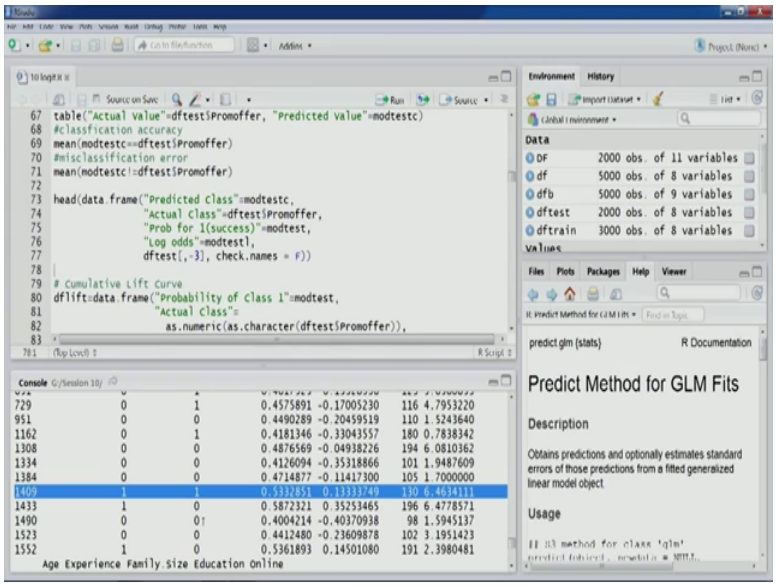
So, now let us look at the some of these observations we can see the probabilities values are co close to 0.5 and Logit values are close to 0, right and all if we look at the weather these observations have been correctly classified you can see first one first row incorrectly classified second third fourth incorrectly classified with the fourth row where we see the correct classification and if we look at a look further then this one is incorrectly classified.

(Refer Slide Time: 17:30)



So, you would see; most of the observation within this range which have probability value in this range have been incorrectly classified right. So, very few observation this is another observation which has been correctly classified.

(Refer Slide Time: 17:46)

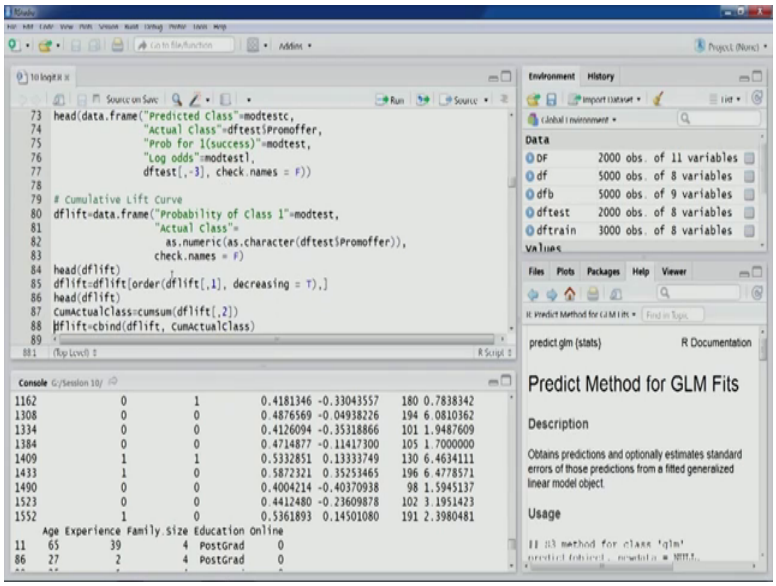


So, very few observations seem to be out of the twenty observation within this range 0.4 to 0.6 that we have seen.

So, in a sense from this kind of analysis we can see that our model is you know our model is able to correctly classify the clear case records and when the situation comes a bit close where the probabilities values are quite close to 0.5 our logit values are close to 0 in those situations the performance of the model performance of the model goes down most of the values are being incorrectly classified; however, if we look at the overall picture the model is giving us 95 percent accuracy. So, that is mainly because of some of the easy some of the direct maybe more observation which are easier to predict.

So, in some situations in this kind of situation we would require expert knowledge. So, the observations which have probability value close to 0.5. So, a in these cases can be identified and you know closure is scrutiny with the help of experts can be done to classify these observations.

(Refer Slide Time: 19:05)



Now a let us look at the cumulative lift curve for this particular exercise for this particular model. So, for this as we have done in a some of the techniques in previous lectures as well. So, I will have will clear this particular data frame first column would be the probability of class one in this case, mod test is storing that information a actual class in this fashion because you can see the code is just slightly you know adjusted. So, that we get the values in numeric form because later on we would be computing the cumulative actual class number. So, you can see promotional offer it was converted into a factor variable; however, the labels where is to 0 and 1.

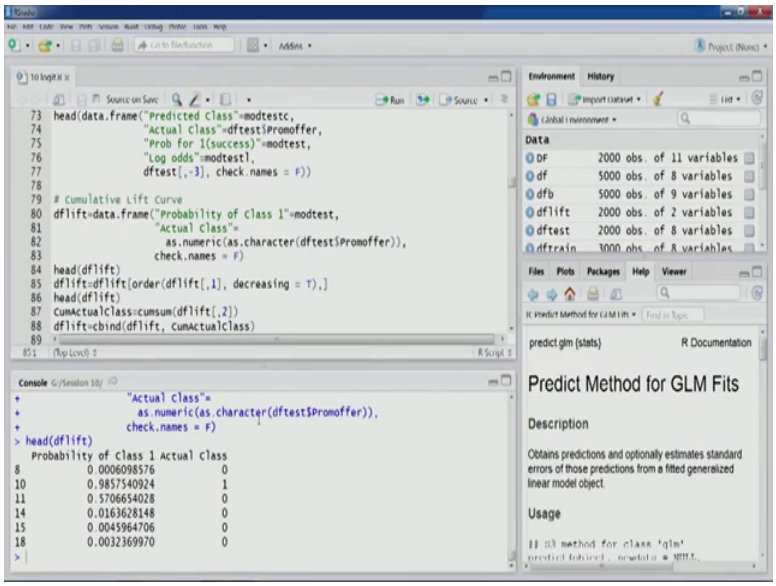
So, we would we would first required to change it to a character variable now the levels would be now the ones. So, labels would be gone and the values would be in correct format 0 and one and then from that we can convert into a numeric format 0 and one right. So, directly the direct conversion factor to numeric might lead to some errors and the values might not be in the desired format. So, if we directly convert from factor variable to numeric variable.

So, the classes would be classes a number of the numeric code for class 0 could can become one and numeric code for class one can become two; however, you would like to have numeric code for class 0 at 0 and numeric code for class one as one because we require certain computation based on that a those values. So, this code will give us the

desired value. So, factor labels for 0 and one and when we converted it into a numeric vector then the values will also be 0 and one using this particular code.

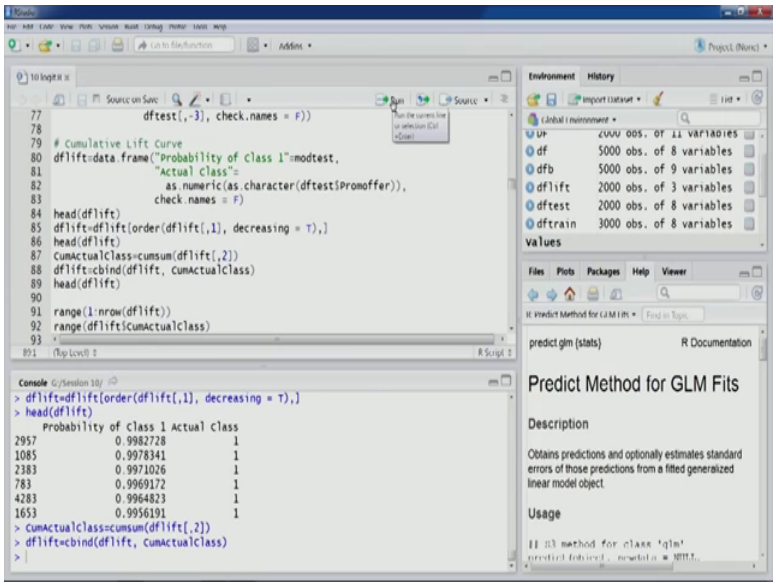
So, let us create this data frame, let us look at the first 6 observations.

(Refer Slide Time: 21:05)



So, we can see in the first column we have the probabilities values and we also have the corresponding actual class. So, please note this that these are the estimated probabilities and the actual class. So, this particular cumulated class in this exercise we have gone through before as well. So, now, what the next thing that we would do we would sort this particular data frame in the decreasing order of probabilities values, right. So, order is order function can be used and the decreasing argument has to be set as true. So, that we get the values in the decreasing order ah. So, let us a run this code let us look at the observations.

(Refer Slide Time: 21:44)

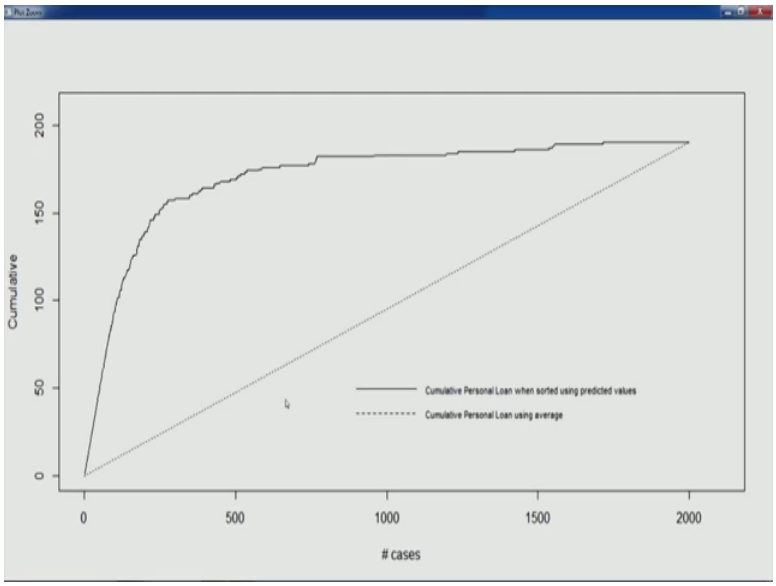


Now if you see that very first row is having the highest probability value followed by the observation with second highest probability value and you would see that first if your observation are all are close to 1.99 something numbers and the actual class is also 1. Now with this with this transformation of this data frame we can go ahead and compute the cumulative actual class ah. So, a this come sum is the function that can be used to ah perform, this computation in our environment ah. So, you can see that in the second column we are applying function. So, we will get the cumulative number in this and stored in this worker variable come cumulative come actual class. So, let us compute this let us add this particular variable in to data frame; let us look at the first 6 observation.

So, now you can see probability and the actual class and the cumulative actual class you can see the numbers also one two 3 four five six. So, now, let us plot our cumulative lift curve. So, first let us look at the range for x axis. So, one to 2000 that is the number of observation in test partition and let us look at the range for a y axis that is range for cumulative actual class. So, 1 to 19 so, that is the range. So, in that sense we can also understand that we have hundred and nineteen in our data set of 2000, we have 190 observations belonging to class 1. So, that is also clear from that.

So, now let us plot you can see that limits x limit y limits are appropriately specified. So, that we focus mainly on the data points the plot region let us generate this plot. So, this is the plot; let us also create the reference line and a legend for the same.

(Refer Slide Time: 23:47)



Let us look at the plot. So, this is our cumulative lift curve. So, as you can see that as we have talked about in when we generated cumulative lift curve for some other techniques, alright so, when we look to identify a first few observations most probable ones right. So, from this from this particular plot we can understand the ability of the model in identifying the most probable ones. So, this lift from the reference line this indicates. So, this particular line solid line is representing the model and the dotted line is representing the reference case different scenario baseline scenario.

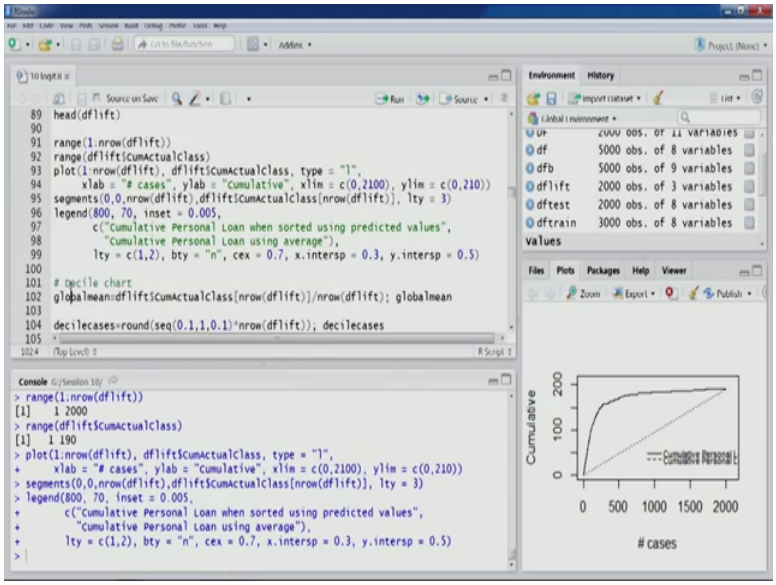
New model new rule and from this we can say our in terms of identifying the most probable ones in terms of identifying the customers who are more likely to accept the promotional offer this model does a good job and provides a provides a good very good lift in comparison to reference in comparison to the benchmark case. So, we can see that lift is quite high in the initial part of the curve and as we look to identify more such cases and the lift keeps you know a lift starts decreasing, right that is because there are just 190 total observations which fall which total observation which actually fall in that category the individuals who have a customer who have accepted the offer.

So, as we go about reaching that number you can see here it is 2000. So, this particular mark is one ninety. So, as we go about a reaching this number the performance of the model start a merging with the performance of new rule; however, in terms of identifying the most probable ones, right ah. So, what we are looking here is the top left corner; so,

this particular corner. So, if we are looking to if we are you know if we are with identifying these many observations. So, the model gives us a quite good performance and comparison to the new rule you can see even at this point we will be able to identify about 150; 100 you know 55 more of more than 150 a individuals you know who who are more likely to accept the offer which is quite close to 190.

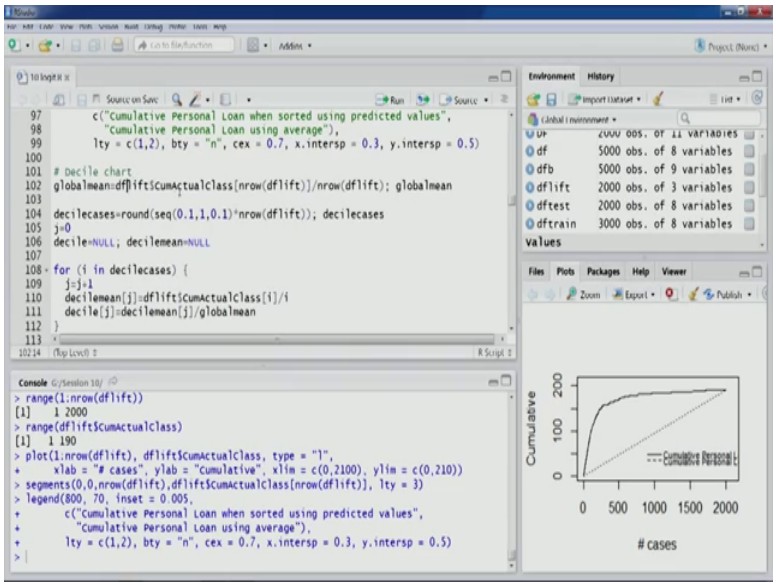
So, in terms of that in terms of identifying the most probable once the model does quite a good job. .

(Refer Slide Time: 26:30)



Now the same information can be further understood using the Decile chart. So, as we have done in previous techniques also.

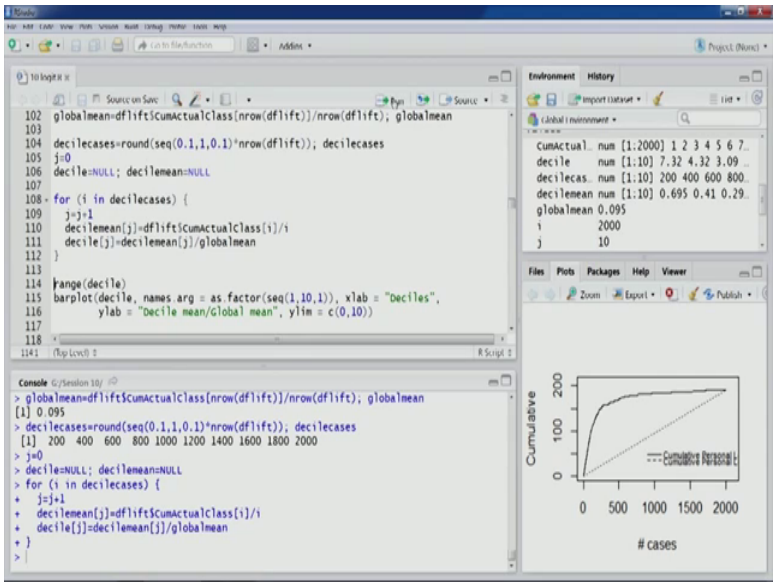
(Refer Slide Time: 26:35)



So, in Decile chart, we will a first have to compute this a global mean. So, you can see that cumulative actual class variable and we are trying to compute the global mean the corresponding value for the last observation and then total number of observation. So, that will give us the, a global mean. So, this is the number point 0 nine five then a Decile cases we would like to have 10 Decile. So, each Decile which represent traditional 10 percent of cases; so, first Decile would represent 10 percent second Decile 20 percent cases third Decile 20 percent cases.

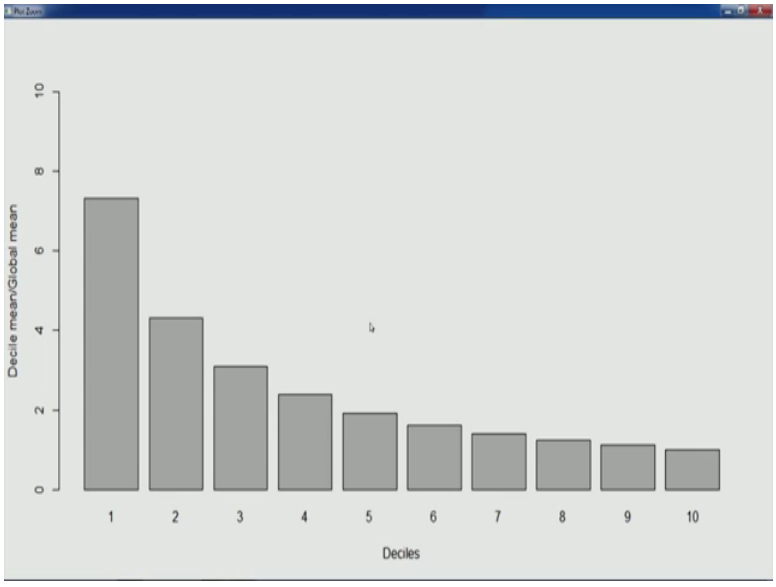
So, in this fashion, we can compute you can see this particular sequence is multiplied with number of observations. So, this will give us the appropriate number of observations for each Decile. So, once this is done we need a counter this counter is basically for the Decile. So, this is actual Decile counter for Decile ah. So, let us initialize this then we have Decile you know a this variable to store the ratio of Decile mean to global mean and the Decile meaning actually mean for each of the Decile. So, let us initialize these variables and in the for loop as you can see this is running from all the values that are in Decile case cases right. So, 10 Deciles and the number of cases and those respective Deciles and once we run this, we will have the numbers let us look at the range of Decile.

(Refer Slide Time: 28:13)



So, this is one to 7.3 something and so, the you can see the limits on y axis have been appropriately specified add 0 to 10, you can also see that other arguments are also for example, on x axis labels that is 1 to 10 Decile; 1 to 10 and other things are appropriately specified. So, let us create the Decile chart.

(Refer Slide Time: 28:42)



So, this is the Decile chart which can be created using the function like we did using a function bar plot. So, we can see. So, a then formation that we saw in the cumulative lift curve the same information is being defective a being depicted in a different format the

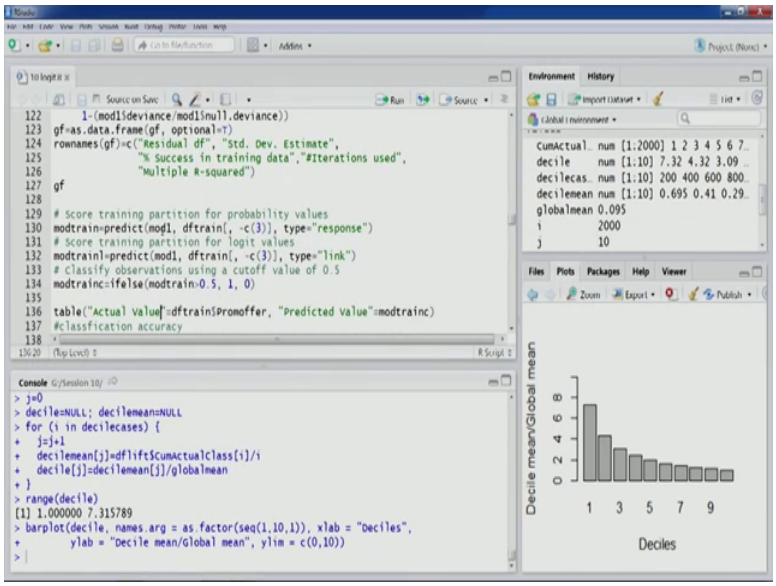
bar chart format in the design chart. So, you can see Deciles. So, each Decile as I talked about first Decile is representing the first 10 percent values.

So, second and the twenty percent thirty percent fashion. So, in a in a way first Decile is giving a a telling us in terms of Decile means y axis a Decile mean divided by global mean. So, this positive cell is giving us the idea about; how well the model will perform in comparison to average case in identifying the most probable ones most likely customers the customer which are most likely to respond most likely to accept the promotional offer. So, you can see that for first 10 percentage a 10 percent of cases the lift is quite high its more than 7, if we look for 20 percent first twenty percent cases and the model still gives us good lift more than a 4 and if we look at the first 30 percent cases the model still gives us good lift more than 2 near about 3 and in this fashion as as we can see just like the cumulative lift curve.

As we look to identify a more number of customer which are likely to accept the offer our lift value goes down the same is reflected in Decile chart if we look for this Decile 4 5 6 7; that means, we are looking to identify most probable 40 percent, 50 percent, 60 percent you know cases. So, our lift will go down. So, typically the, you know; we can we can go for the up to or Decile where the lift value is still greater than one. So, we look at near about you know a eighth Decile; that means, a 80 percent of the cases, this is about near about seems to be near about one. So, from this also we can understand that out of 190 observe of 190 observation which have you know 190 customers which have accepted the promotional offer about 80 percent of them can be easily identified by the model.

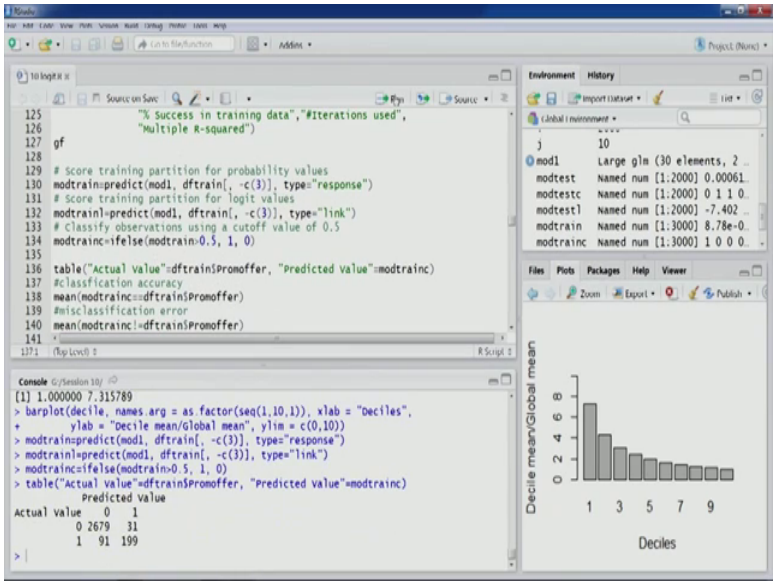
We can further look at some of the measures of goodness of it. So, these are some of the values.

(Refer Slide Time: 31:20)



So, we will discuss some of these values and later lecture now, right. Now what we will do will look for our performance in the training partition. So, the performance that we have seen till now is watch for the test partition now do the let us do the same exercise on training partition itself. So, let us have a look on the same. So, let us compute the probabilities values followed by Logit values and followed by classification just like we did for test partition. So, let us look at the classification matrix.

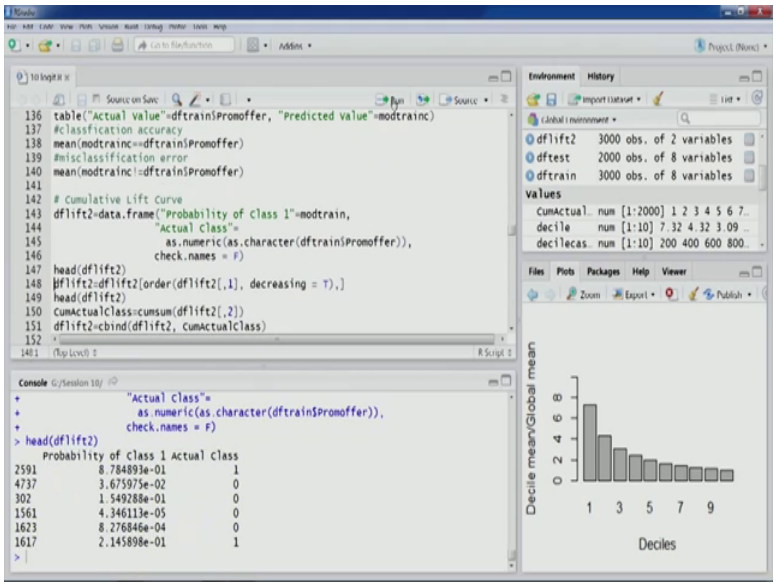
(Refer Slide Time: 31:51)



Here we can see out of 3000 observation good number of observation majority of the observation have been correctly classified as we can see in the diagonal elements, right.

So, now let us look at the classification accuracy you can see 0.959. So, this is more than the performance on tests partition which is expected because these are the observation on which the model has been built. So, the error is this much about 4, we can also create cumulative lift curve and Decile chart for this particular partition as well.

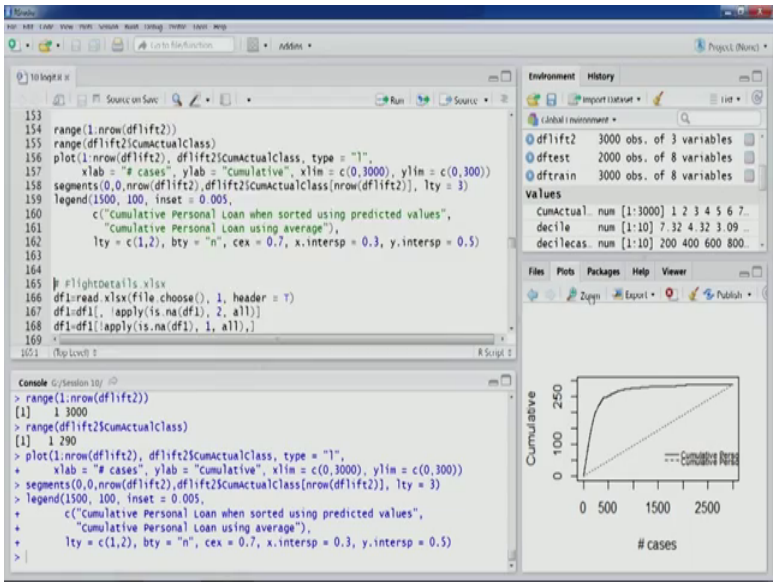
(Refer Slide Time: 32:34)



So, this data frame is created and let us look at these first 6 observation. Now, we let us sort it out let us order it in the decreasing value this is.

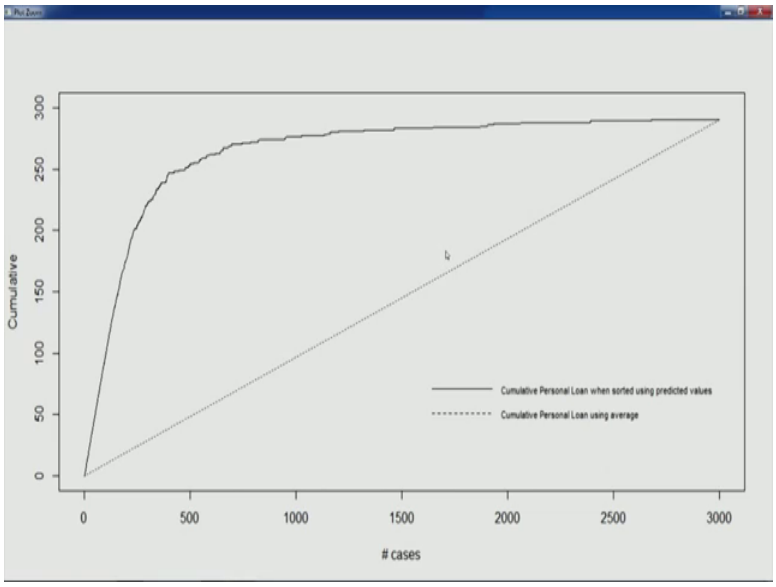
So, most of the values we can see now first 6 observation close to one let us compute the cumulative values let us look at this. So, a once this is done we can go ahead and a create our lift curve.

(Refer Slide Time: 33:04)



So, we can see here. So, this is the curve for the training partition.

(Refer Slide Time: 33:09)



So, we can see because the model is doing good on test partition also. So, both training and test lift curve look a quite similar. So, with this will a stop here and we will do another exercise to understand further non logistic model in next lecture.

Thank you.