Business Analytics & Data Mining Modeling Using R

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

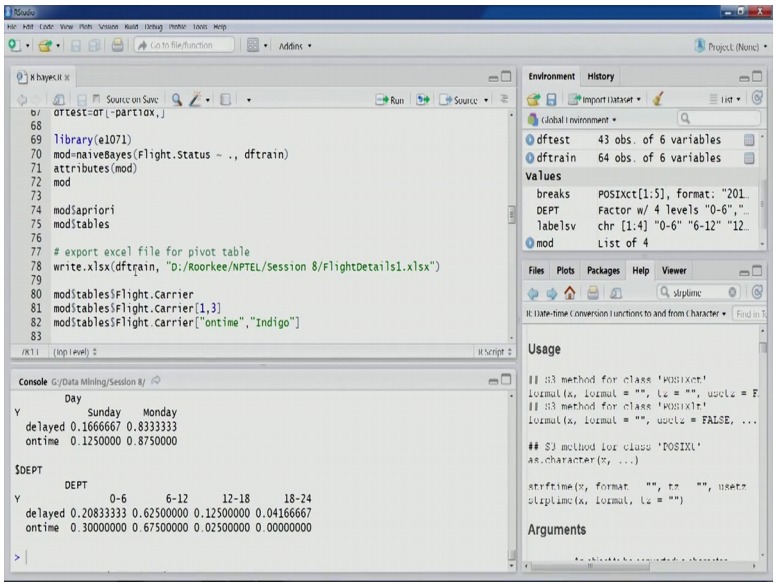
Lecture - 35
Naive Bayes - Part V

Welcome to the Business Analytics and Data Mining Modeling Using R. So, in the previous lecture we were discussing Naive Bayes and specifically we were doing an exercise and modeling exercise in R environment. So, we were able to import the data set and then the process and some of the transformation that we did in the previous lecture. We also looked at some of the issues that we encounter related to dates and arrival and departure time and how we were able to correct those dates that come because of the importing of data from an excel windows, PC to R environment.

And then later on be process to be created the different time intervals based on the departure time, actual departure time. We also did our mod exercise the partitioning and modeling. So, let us, we also looked at the tables the conditional probabilities values for different you know combination outcome variable and predictor combination for different values that is nothing but different categories for the predictors and the outcome variables. So, we looked at all those things.

Now, some of these things can also be performed using a pivot table and excel some of these conditional probabilities. So, though we rely on the computation in R using this particular functions, but these are essentially the mathematical computation computation essentially the technique is more of mathematical in nature and the probability computational or all the crux of this Naive Bayes modeling. So, what we will do? We will export the data set, and that for the training partition that we have done we have created this training partition that we have created in the excel format. So, we will also learn how to export data into excel format.

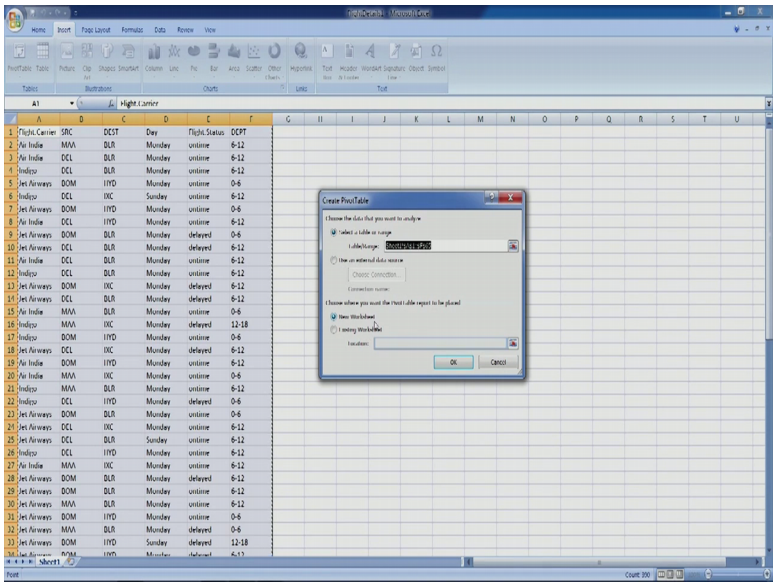
(Refer Slide Time: 02:12)



So, and then we will do a pivot table we will create a pivot table and see how the same conditional probabilities that we have computed using a Naive Bayes function that can also be done using pivot table.

So, let us find and approve it folder for this first. So, if this is the folder then we can specify in this name here and as we have discussed before that we need forward slashes and not the backward slashes in R environment to be able to use the absolute path or files. So, right dot x l s x is the function if we want to export the data into an excel file. So, the training partition data is going to be exported here now we execute this line it has been processed and a file flight details one has been created.

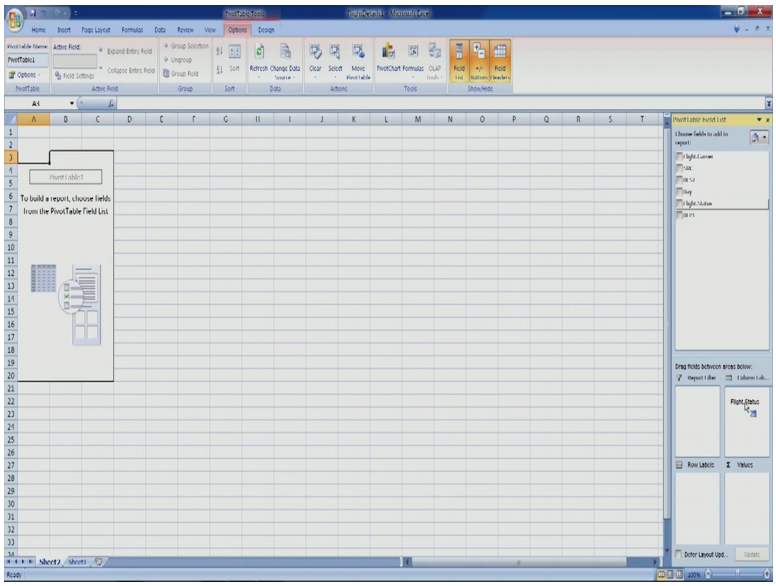
(Refer Slide Time: 03:37)



So, let us open this. So, this is the file you can see first, first column is nothing, but the serial numbers. So, nothing but the index is also these are indices for which were selected when we created the partition. So, right now we do not need this further, column. So, you will delete this and then we have the other columns we have the 6 variables that we require for our modeling exercise. So, these are the 6 variables as you can see here.

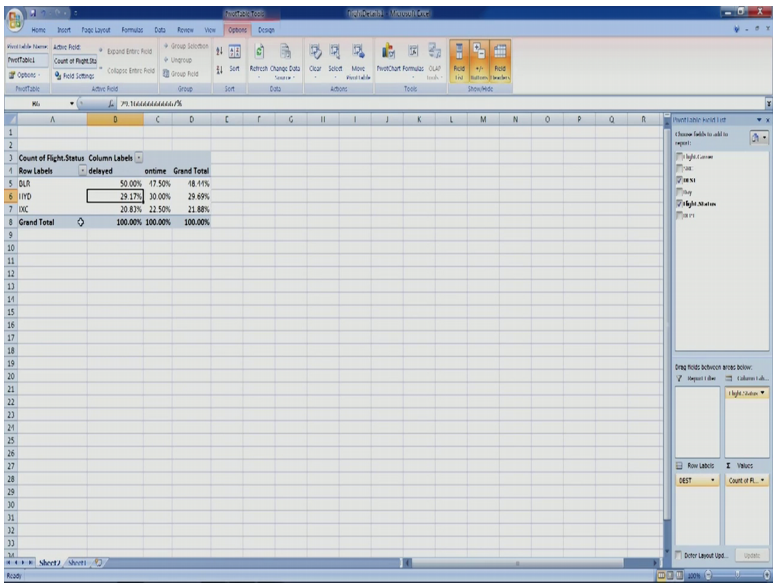
Now, to create the pivot table will select the data set and including the header. So, we will select using ctrl shift and down arrow. So, once this is done we can go to the insert tab within excel and then we can see that pivot table option there and then here we can click on this and we will get the drop down menu and first option is pivot table. So, let us create this. So, the range is already selected.

(Refer Slide Time: 04:53)



Now, we want this pivot table to created on a new worksheet, so we will just do and this is what we get. Now, from here we can create pivot table for different pairs of variables. So, one is flight status this definitely has to go here and column and then for example, depending on the destination or example you want to calculate values for destination and so that we can do.

(Refer Slide Time: 05:14)



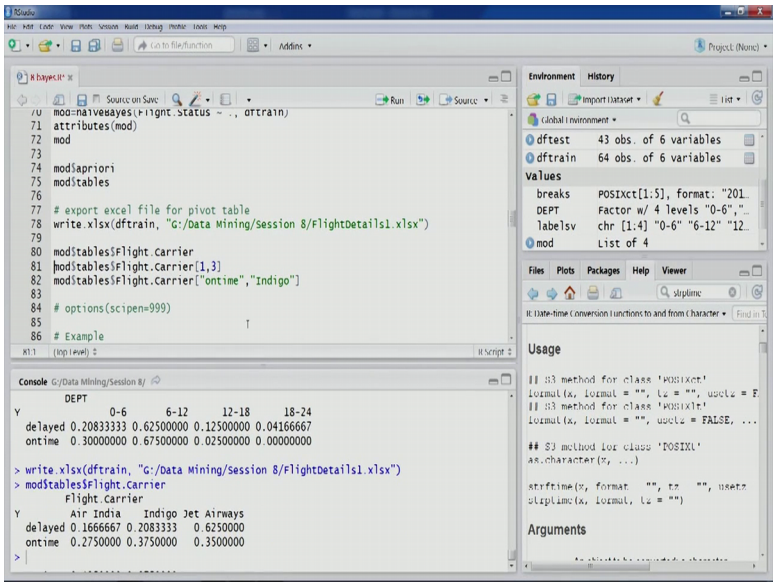
So, the column label is flight status row level is destination as you can see a particular pivot table has been created here you can see and the count of count is the default and

has been taken here for flight status few things will change. So, for this left click you would see value field setting and the value field setting. So, everything is ok, and this is summarized by is everything, but show value as there we would like to change and we would like to make it percentage of column. So, this as you can see once we did this. So, the numbers have been converted into percentages.

Now, if we can compare this, to the results that we have got in R to find out whether everything is ok or not, you can see let us look at the results that we had for output a variable and the destination. So, you can see 3 categories b l r, h y d and IXC. So, numbers you can see here 50, 29.17, 20.83 same numbers are here, similarly for on time 47, 47.5, then 30 and then 22.5. So, you can see the probability computation can also be performed using just we just looked at the data created a pivot table and we were able to compute these conditional probabilities. So, the computations that we have performed using the function in R this can actually be performed using a pivot table that are available in excel as well.

So, let us move forward. Now, the table conditional probability table if we are interested in accessing specific values for example, in the table if we just want to have a look at the flight carrier probabilities, conditional probabilities related to flight carrier, we can express in this format because the tables that we have is actually a data frame a list and we can execute this you can see that for a out for outcome variable and the flight carrier we have the conditional probabilities.

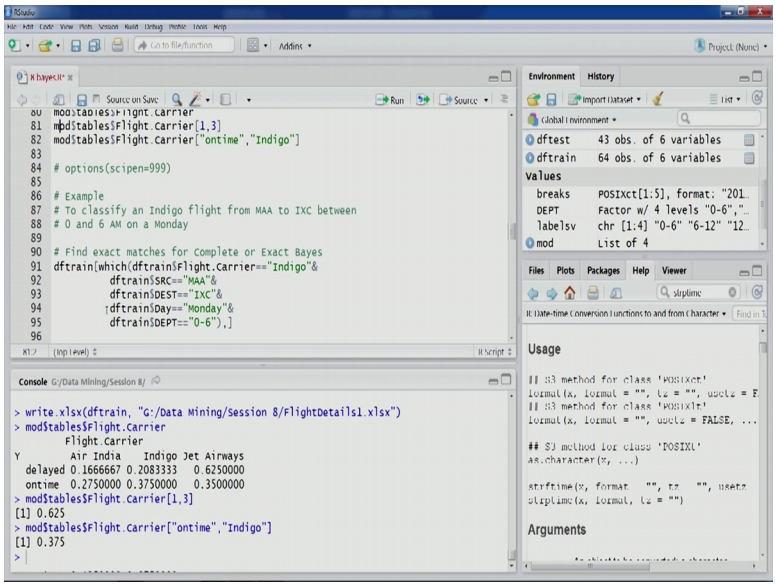
(Refer Slide Time: 07:45)



If we want just to one first row and third value that is the value for jet airways, so this is how we can you use the brackets notation and axis this value you can see 0.625, if we do not want to use the numbers for row and column we can mention the name of the name of the rows and columns also in this fashion on time and indigo or second example, will get this value 0.375. So, let us execute this you can see. So, this is how we can access the values in R for the Naive Bayes tables from the Naive Bayes tables.

So, let us go through an example where we will try to compute the values by accessing these numbers.

(Refer Slide Time: 08:46)

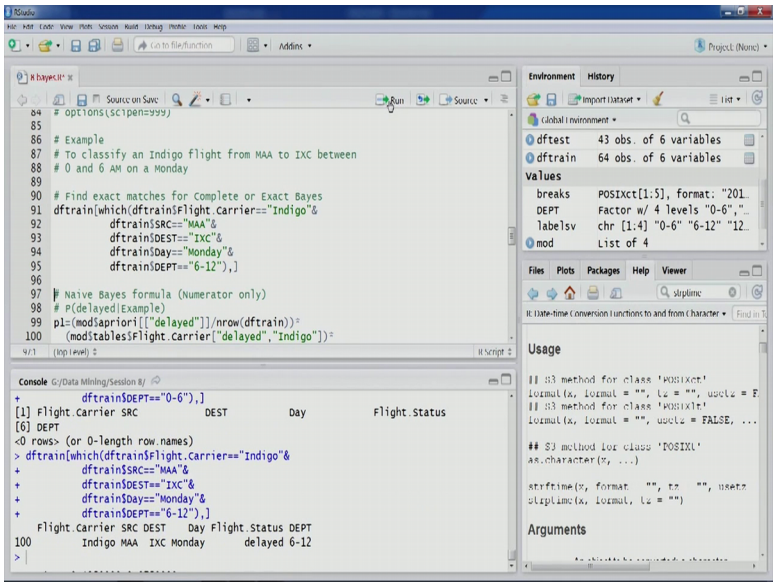


So, probability numbers as we have shown we can also be computed using excel and R as well we can access the individual probabilities value. Now, using them using these value we can now go through one example where in we will try to compute the probability values. So, this is an example.

We want to classify an indigo flight from MAA this particular airport to IXC this particular airport between 0 and 6 am on a Monday. So, you can see in this example. So, all the information related to different predictors is available. Now, using this predictors this information how we can go ahead. As you know that in complete exact way this will have to find the exact matches, which we are not going we are going to use the Naive Bayes computations.

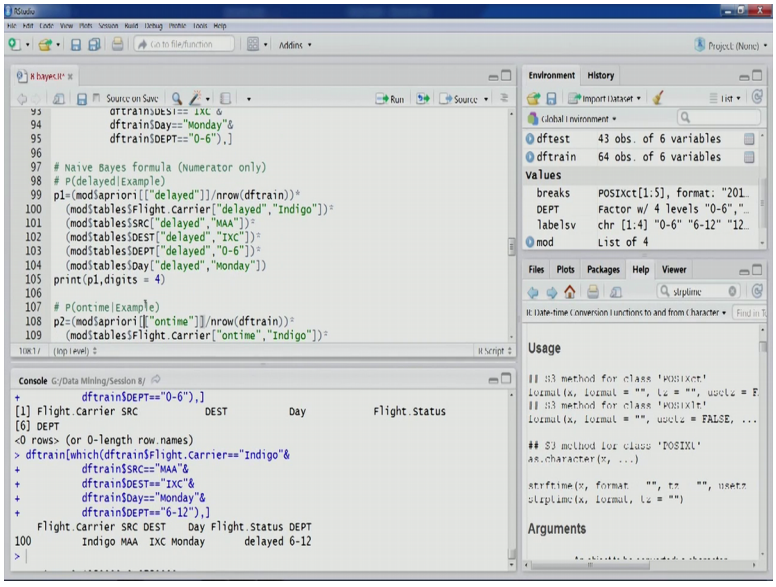
First let us see whether there are any exact matches or not. So, we will just look at whether there are exact matches or not. So, this is how we can do this in the d f train partition and here we can look at whether the flight dot carrier is having this value indigo and that in the source and destination and then day and departure. So, appropriately we can mention the value and find out the rows you can see this particular expression has been written in the row value right, the column value all columns are selected. So, let us execute this, you can see 0 rows. So, no such observation is there in the training partition right. If we just changed this particular value from 0 6 to 6 and 12 let us see if there are any values.

(Refer Slide Time: 10:41)



So, we just change the interval you can see there is one value which is actually having all these. So, there are other predictors information or net we have a match for other pattern information except for the except for the departure time interval. So, for that particular interval we did not have any matches therefore, compute exact this cannot be applied.

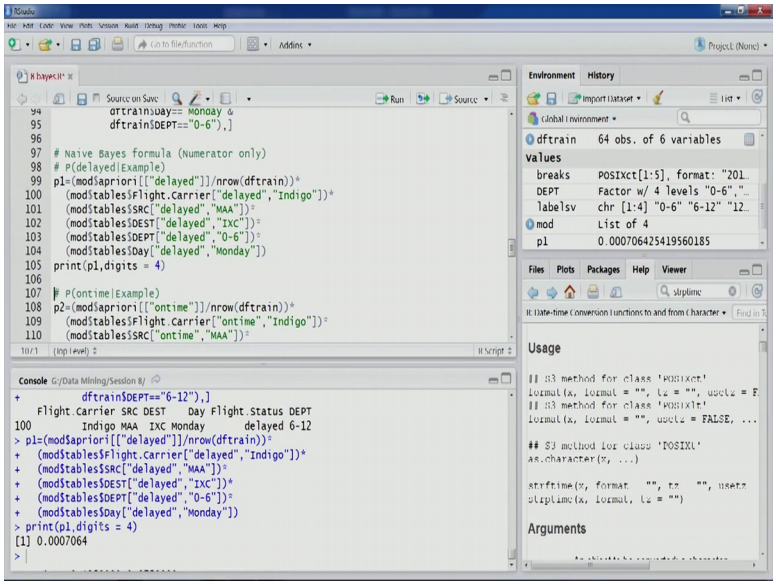
(Refer Slide Time: 11:14)



So, let us come to the Naive Bayes formula. So, we are going to compute numerator value and numerator value only first. So, as we have discussed in excel exercise as well first we need to compute the probability value for delayed given the information given

the predictor information as per the example. So, you can see is how we can access. So, first a proportion of values belonging to delayed class. So, this is their then delayed and indigo and delayed MAA and delayed IXC delayed 0 to 6 time interval departure time interval and then the later on Monday. So, for all these in this fashion we can actually access the individual condition probability values and we can multiply and the proportion is all is also there. So, this is how we can actually compute the value you can see p 1 has been computed.

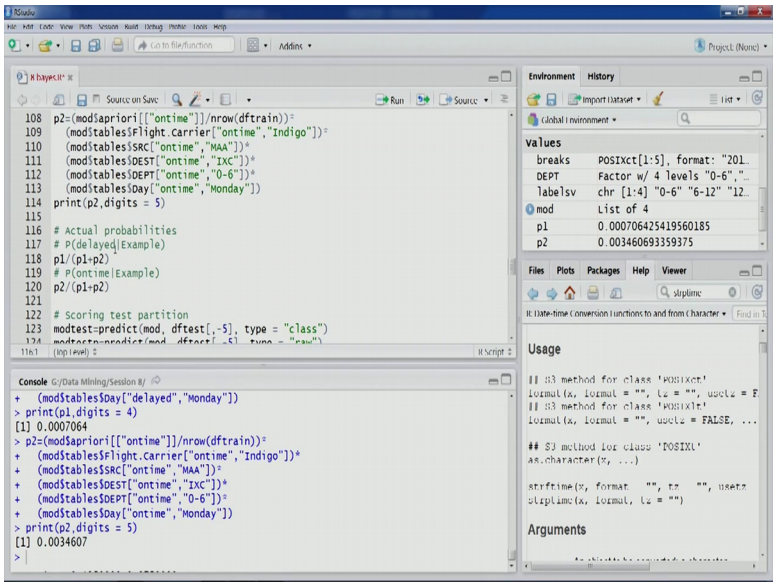
(Refer Slide Time: 12:06)



So, let us print up two values in different digits. So, this is you will get 0.0007064 this is the probability value. Now, for the on time class also we can perform the similar computation here also as you can see that mod and then you can see on time proportion of records are belonging to the on time class. So, that is there.

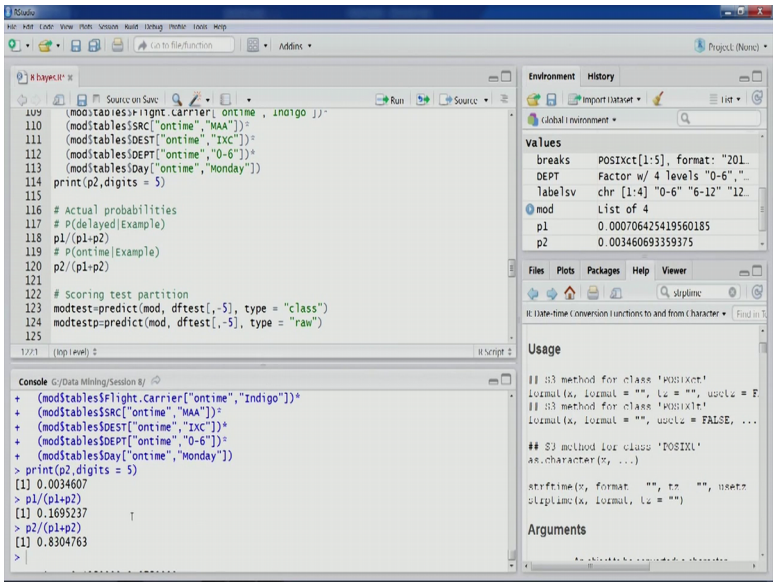
Then for flight carrier you can see on time and indigo, then for source you can see one time and the MAA this particular airport and then the one on time IXC this the destination airport then the time interval departure time interval this is for on time 0 to 6 and the last value that is for Monday. So, let us compute this. Let us also open the value up to five significant digits. So, now, these are the values the for the numerator part.

(Refer Slide Time: 13:09)



So, as we have done in, as we have discussed before also that we want to compute the actual probability value we can do so, using this p 1 we can divide by the summation of these two value we have just computed. So, we will get the actual probability value.

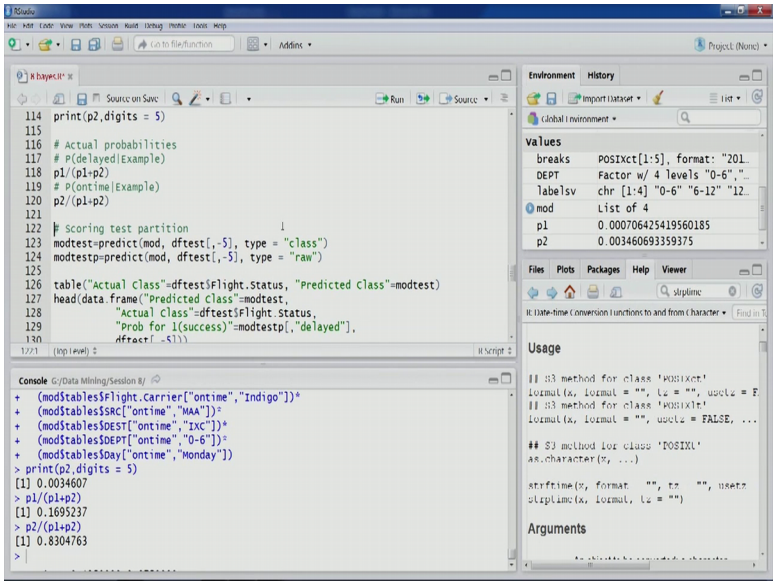
(Refer Slide Time: 13:20)



So, these are the actual probabilities value. Now, once we have the probabilities value depending on whether we want to use the most probable class method or depending on whether we want to use the you know we have a class of interest. So, then in that case we

would like to specify the cut off value and compare using that, based on that we can always classify the observation.

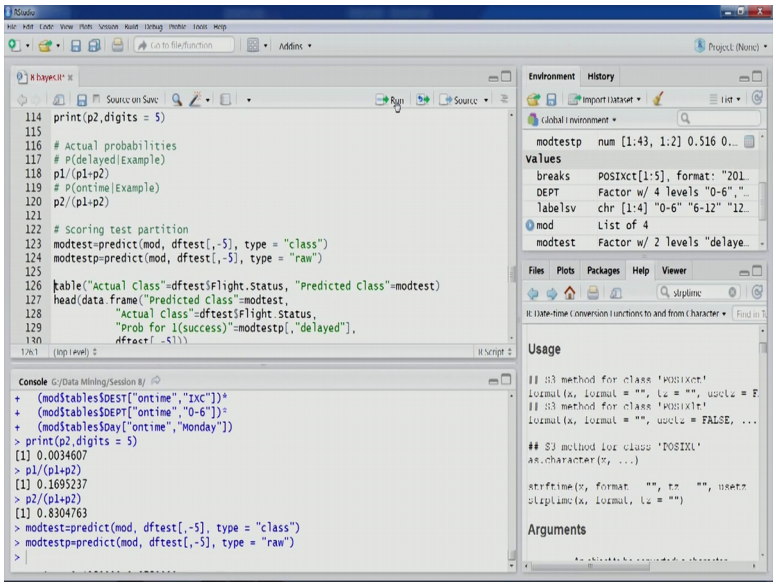
(Refer Slide Time: 13:52)



Now, once we are done with our building our model using the training partition. So, let us score our test partition and evaluate the performance of this Naive Bayes model on the same. So, what we are going to do here is use the predict function as we have been doing in previous techniques as well and let us say score this.

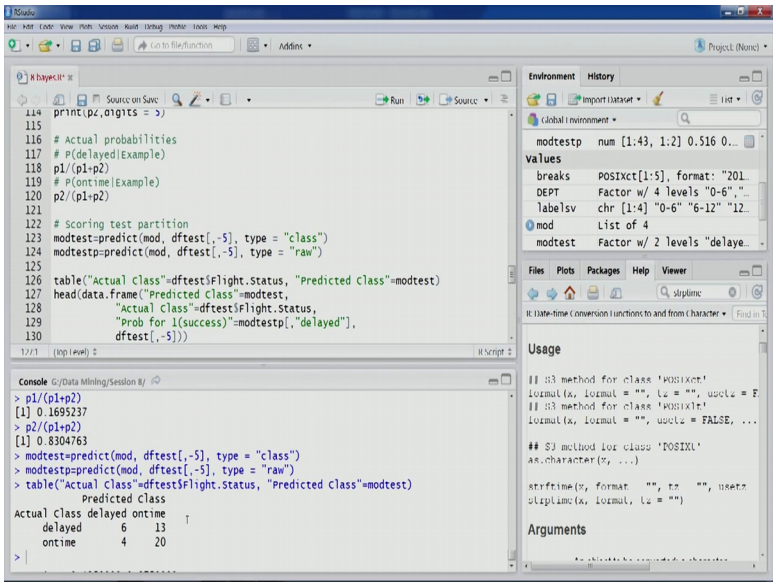
So, we have two options here type is the argument that can be used within the predict function. So, if we use type as class then we will get the, we will get the scoring for the actual class and if we use this particular time as raw then we will get the estimated probabilities value. So, let us compute each of these two to actual class membership and also the estimated probabilities value.

(Refer Slide Time: 14:41)



So, let us look at the classification tables that we can generate using table function. So, you can see actual class stored in the flight dot status for test partition as well and the predicted last just be now just now we computed.

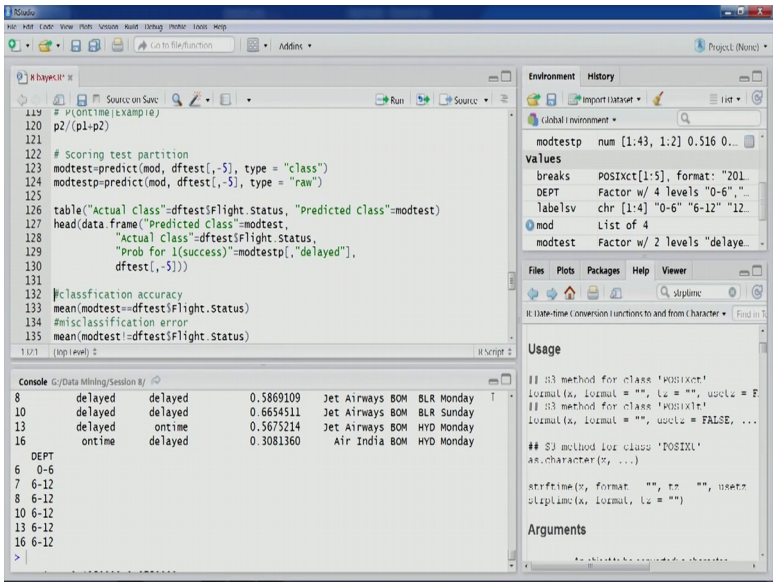
(Refer Slide Time: 15:01)



So, these are the results as you can see from this particular classification matrix that 6 delayed flights have been correctly classified as delayed, but 13 of them has been incorrectly classified. On time 4 of them have been incorrectly classified then, but 20 of them have been correctly classified. So, this gives us the idea. So, you can look at 26

observations have been correctly classified and 17 observation have been incorrectly classified this is mainly because of this smaller sample size that is why you are not getting that good result.

(Refer Slide Time: 15:46)



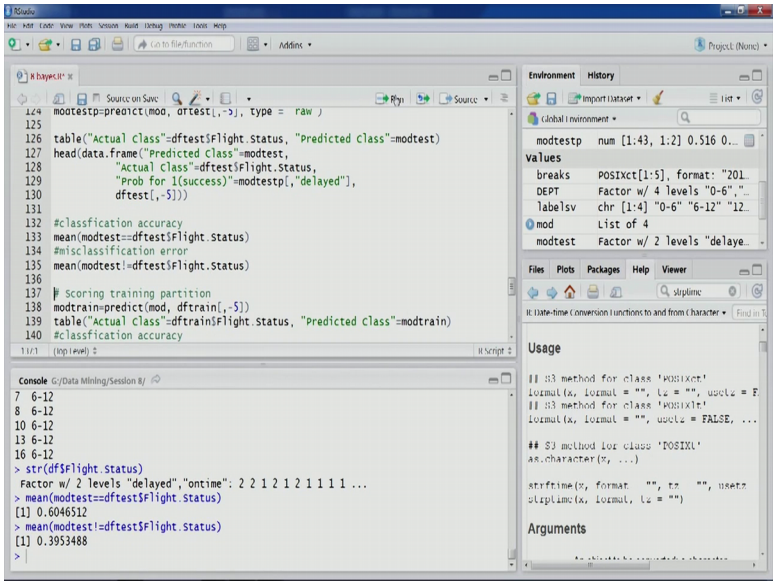
Now, we can also have a look at the full information the predicted class the actual class and other on the probability value and the other variables to understand what has happened in the modeling process. As per the results can see predicted class an actual class you can see that these values the probabilities here again.

Again you can see that in this case the default class for our variable default class for our variable we can look at here let us look at the default class you can see delayed and on time. So, the delayed is the going to be the reference category and on time is the main category included in the model. So, would see these probabilities values that p c here is. So, this is a probability of a particular class belonging to on time category, more than 0.5 value of the value is more than 0.5. So, it will belong to the if the does the probability is such then this is going to has been predicted as delayed if it is more than 0.5 if it is less than that and it has been classified as on time.

So, predicted values also we can see here. So, for any probability value that was more than 0.5 we have classified it as delayed and the probability value of less than 0.5 this one from this we can understand it has been classified as on time. So, if we want to look

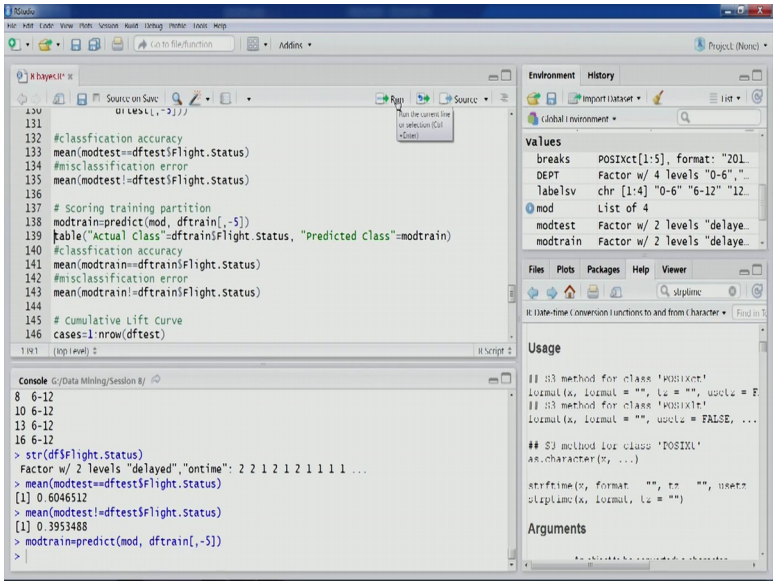
at the actual classification error accuracy and misclassification number this is all we can do it.

(Refer Slide Time: 18:19)



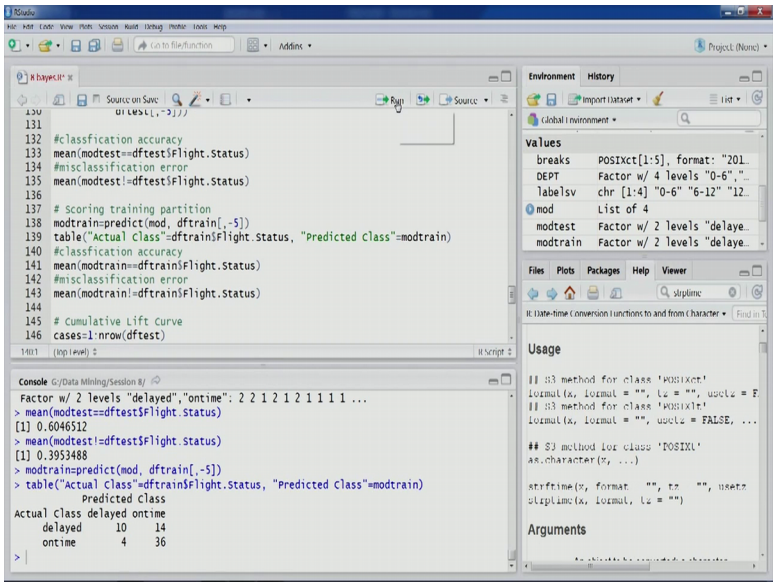
So, a comparison between these cold values and the actual values will give us the accuracy. So, that was 0.6 and other was the remaining 0.39 or about 0.4.

(Refer Slide Time: 18:25)



Now, we can also compare the performance of model on test partition with the performance of model on training partition. So, let us the score the training partition using the model itself. So, the same function predict.

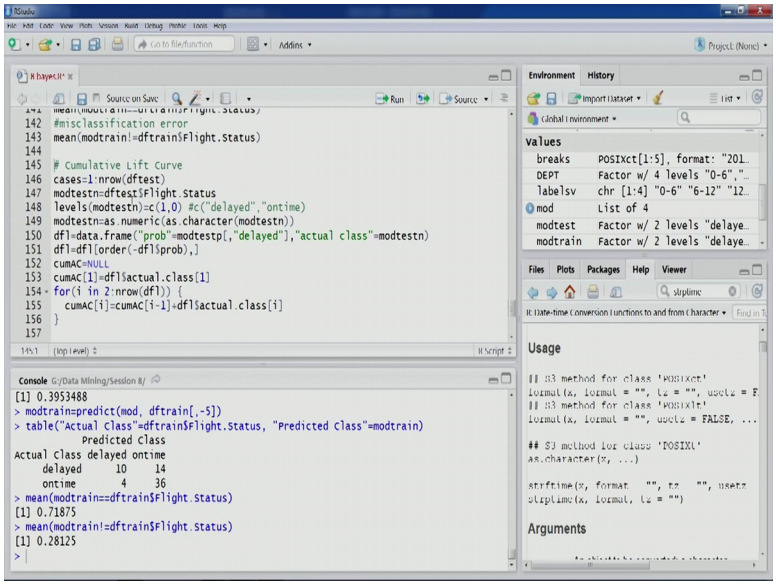
(Refer Slide Time: 18:43)



And we can score this we can look at the classification matrix here you can see many more observations are classified correctly in this case as expressed using the accuracy and error numbers.

So, model in performing much better for the training partition main reason being it was built on training partition itself and perform is performing slightly poorly for the test partition. Now, this is also because of the smaller sample size that we have now for this model if we want to generate lift curve.

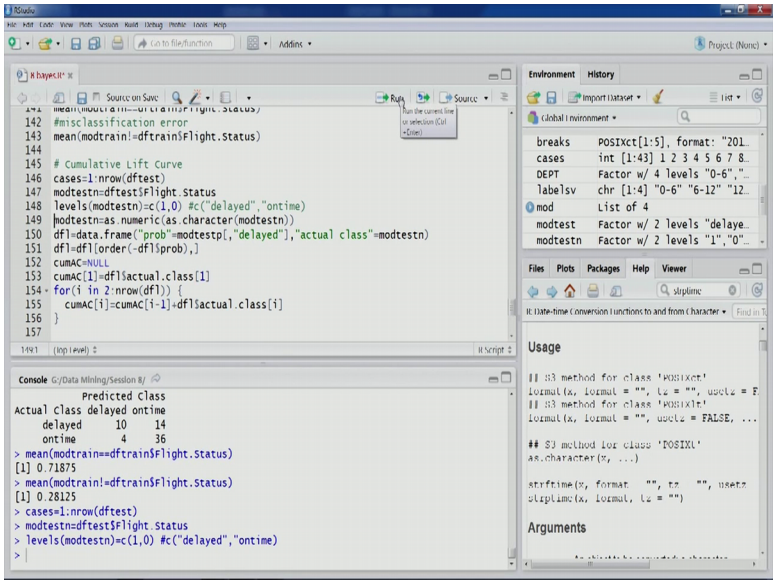
(Refer Slide Time: 19:19)



So, this is how we can do it. So, cases we have how many cases. So, for the test partition and I we want to create generate the lift curve cumulative lift curve lift curve. So, we have the 43 observation. So, let us create this.

Now, the values actual classify, actual values for this particular partition test partition let us access them using more test and variable. Now, labels let us change the labels because we need cumulate we need cumulative values for based on the actual, based on the actual classification of the observation. So, we will like to change the labels. So, in this case delayed and on timed one for delayed and 0 for on time.

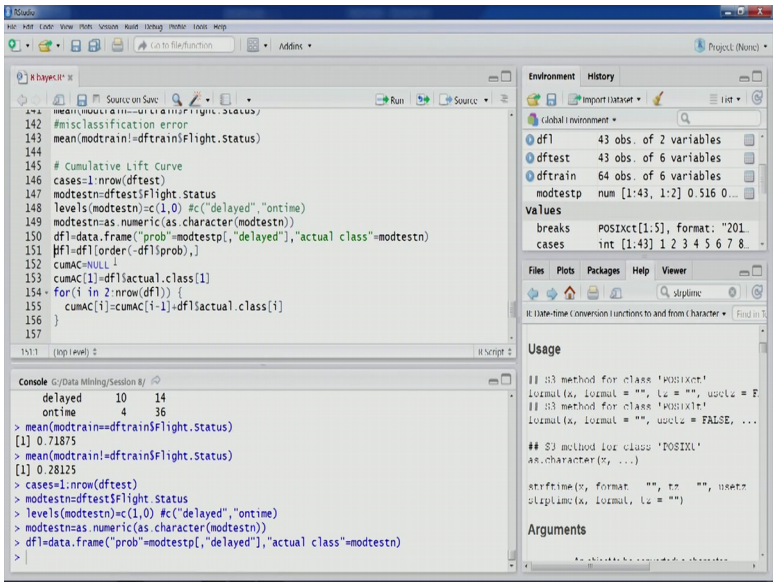
(Refer Slide Time: 20:18)



So, let us change these labels. Once this is done because this mod test and being the factor variable. So, let us convert it into numeric variable so that we can apply the operations that is the mathematical operations. So, this is you can see now numeric variable and 0 and 1. Now, we would be able to compute the cumulative values right.

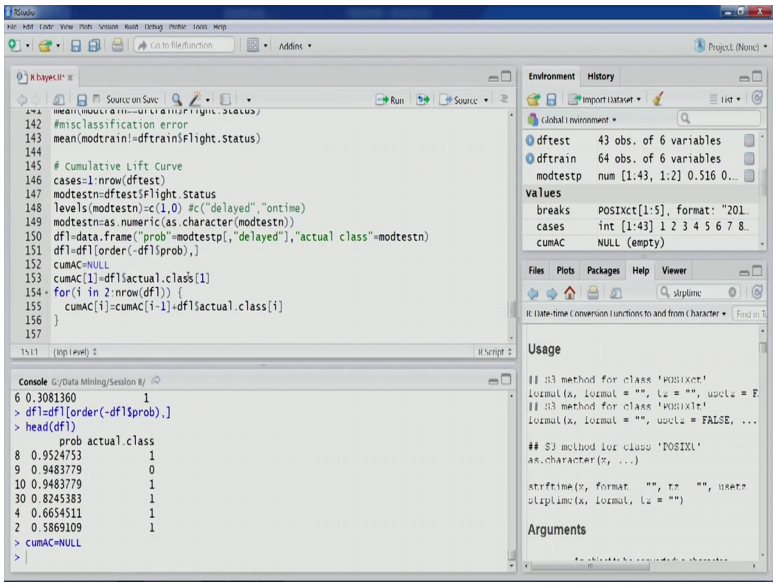
So, using this, using the probabilities value that we have that we have computed earlier right for the delayed class right, using this particular value and the actual class that we have just now created a variable we can create a data frame of these two variables.

(Refer Slide Time: 21:05)



And then we can sort these variables in terms of the probabilities values the decreasing values right. So, we can if you are interested in looking at the few values. Let us look at the first 6 observation in this particular data frame.

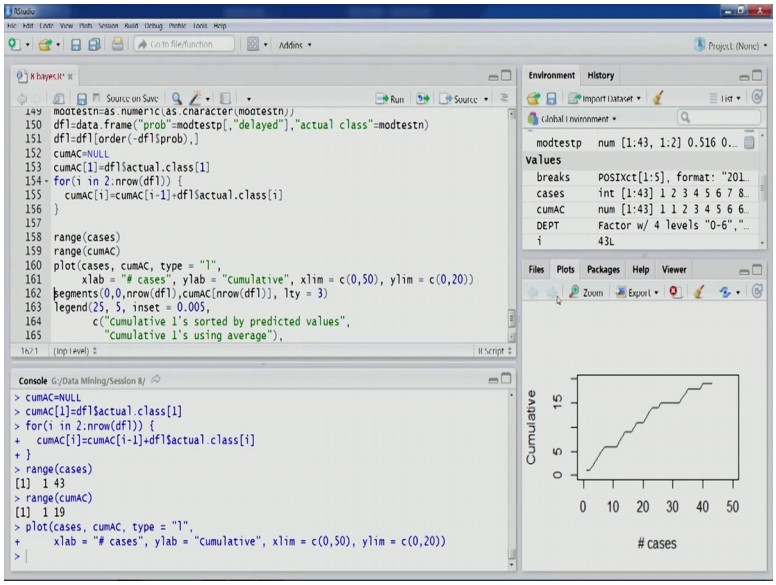
(Refer Slide Time: 21:18)



You can see probabilities value in actual class, let us order them, as we have done in previous lectures at as well. So, you are again interested in looking at few observations you can see now these values have been sorted or ordered in the decreasing probabilities values sequence right. Now, we can compute the cumulative values. So, let us go through

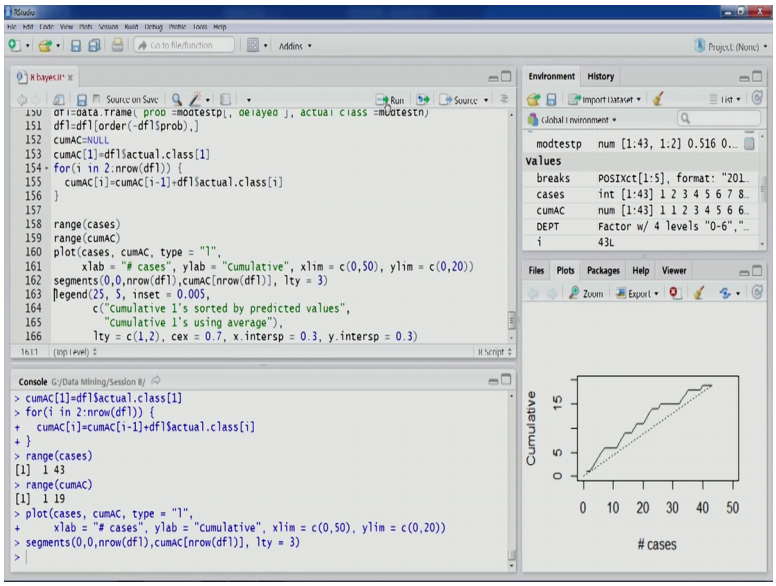
the code for the same. So, the first value is going to be the same. So, how we can do and then for the remaining values we can run this loop when in the previous value is going to be added to get the cumulative value. So, once this is done let us look at the range. So, we have 43 cases, let us look at the cumulative value range 1 to 19 then as you can see appropriately the limits have been specified and other things there.

(Refer Slide Time: 22:19)



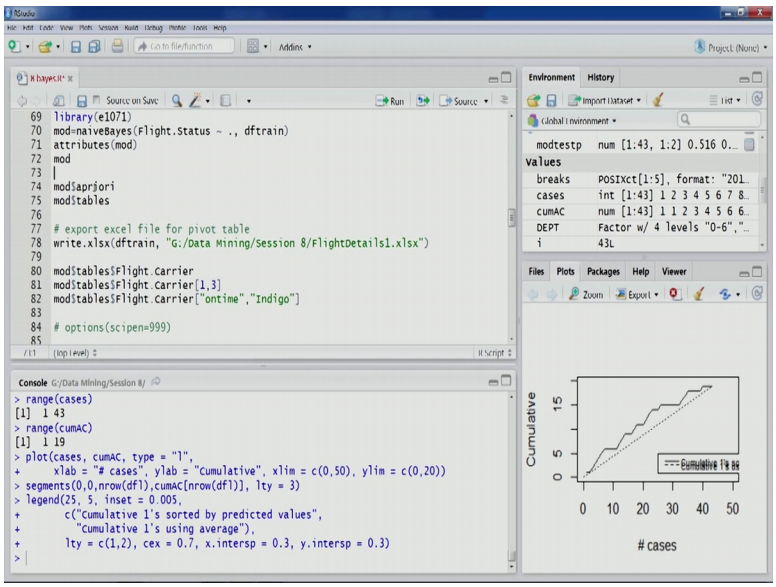
So, we can generate the plot. So, this is the cumulative against curve that we have, this is the plot that we have now if we want to compare its performance with the reference line. So, this is what it is.

(Refer Slide Time: 22:35)



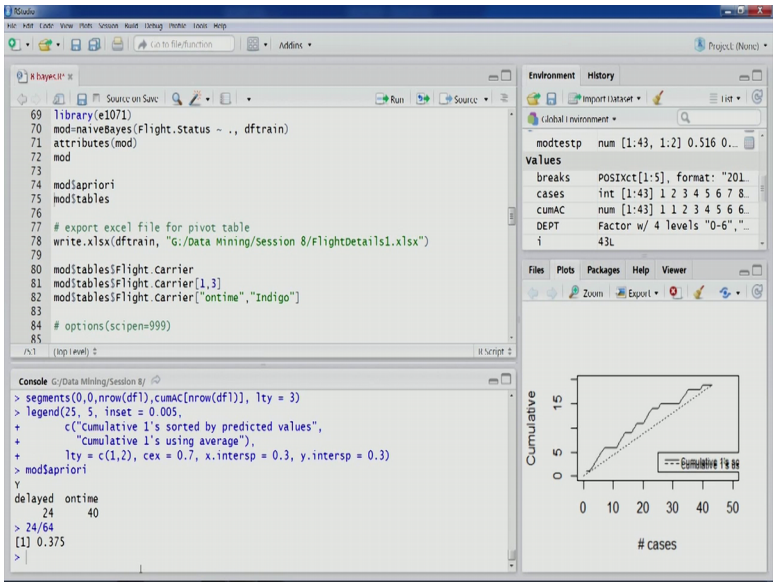
And let us look at the plot. So, you can see this particular plot is performing better than in terms of identifying the most probable ones or most probable delayed flights. So, this model does quite a good job with respect to the average case that is the reference line has so many reference line right. However, if we just look at the prior probabilities and the actual numbers that we have computed right the naive bayes over all on comparing with the probabilities we can find out that if here do the naive rule then what would have been the scenario.

(Refer Slide Time: 23:23)



And if we applied the Naive Bayes modeling, you can see here. So, had we classified all the observations as belonging to the majority class following naive rule then added observation would have been classified as on time and we would have had the error as just 24 divided by 64 and 0.375 right.

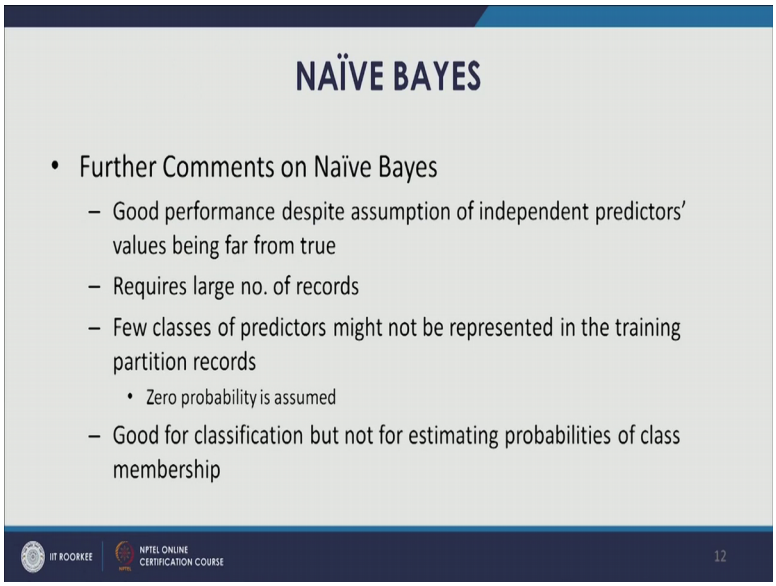
(Refer Slide Time: 23:55)



But if we look at the training partition error that we have computed just you know before this right, it is the training partition is model is doing good and training partition we look at the models performance on test partition you see the error is 39.5 and here you is you would seen naive rule the error is 37.5. So, now, rule is actually doing better than our model, but it is the cumulative lift curve that we saw that is giving the usefulness that is indicating towards the usefulness of the model in terms of identifying the most probable ones most probable delayed flights.

So, this is the our modeling exercise and that we wanted perform let us go back to our discussion on Naive Bayes. So, what we will do? We will discuss Naive Bayes few more points.

(Refer Slide Time: 24:57)



So, we look at the Naive Bayes as a technique its quite simple right quite simple and easy to understand is based on the probabilities values. So, the interpretation is quite easy and can give us surprisingly good results in some situations especially when that independent that predictors independent of predictors values that a junction is actually met then it can give us the surprisingly good results or sometimes even outperform other techniques.

So, this good performance can come as mentioned in the first point despite assumption of independent predictors values being far from crew. So, even if this assumption is not being met even then we get good performance using Naive Bayes. However, it requires a large number of records because we would like to cover as many scenarios as possible because the predictors they are mainly categorical in nature. So, therefore, the different scenarios that could be there we would like to compute most of them. So, that for a new observation there is always probabilities values for to compute and assign a class to that observation. So, requires a large number of records.

Now, if we do not have a larger sample size then few classes or predictors might not be represented in the training partition record. So, in such situation 0 probability is going to be assumed and then it would be difficult to classify that particular observation right. So however, I mean this particular case is cannot be handled by other techniques as well, but the situation is more complicated in naive Bayes main reason being that the information

in other predictors. For example, for one particular out of 4 or 5 predictors it is just to the one predictor that is not matching, and if that is not if that is then, that is not present in the training partition that one particular class then the value 0 probability value is going to be assumed and therefore, all other predictors. So, the remaining predators information would not be counted which is not the case in other techniques.

Far from this Naive Bayes is mainly suitable for classification tasks as we have discussed in previous lecture previous lectures as well. Another important point is good for a classification, but not for estimating probabilities of class membership. So, as we have talked about that in the Naive Bayes formulation it is the numerator that is sufficient for us to get achieve the rank ordering of rank ordering with, rank ordering of different my classes of outcome variables and that is what we are interested in. So, the model as such will do a good job; however, if we are interested in the actual probabilities values. So, then Naive Bayes is not a suitable technique for such scenarios.

So, this will stop our discussion on Naive Bayes. In the next lecture we will start our discussion on classification and regression trees.

Thank you.