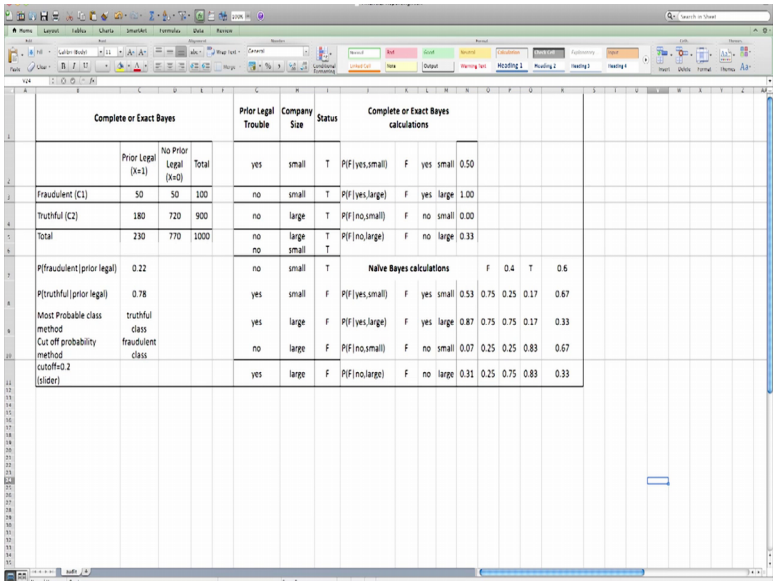
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture - 33
Naive Bayes – Part III

Welcome to the course Business Analytics and Data Mining Modeling Using R. So, in the previous lecture we were discussing in Naive Bayes and specifically we looked at the formula for Naive Bayes. So, specifically we looked at the Naive Bayes formula, we also looked at the a steps for the same when we have a class of interest, when we do not have a particular class of interest the general or typical scenario that we have also discussed in the previous sessions that is. So, these are some of the steps that we have discussed in the previous.

So, now, what we are going to do in this particular lecture, we are going to do a small exercise to understand Naive Bayes and the different and the various steps for Naive Bayes modeling computations and other things through an exercise. So, we will do an exercise using excel and then followed by an exercise in R.

(Refer Slide Time: 01:17)



So, in the previous session we had gone through one example related to complete or exact Bayes. So, you can also see that classification matrix here as well where we had

computed the probabilities right fraudulent given for legal. So, here this particular example was where we have just one predictor and we had computed the probabilities of belonging to that fraudulent class and truthful class given that predictors information and based on that using most probable class method and cut off probability method with cut-off 0.2 also we had seen what would be the assigned class now the same thing can be extended especially using the Naive Bayes calculation if we have you know. So, 2 predictors let us go through an example where we have 2 predictors.

So, first one in this case is prior legal trouble as was in the previous case as well and the second one being the company size now the status of those financial reports whether they were found to be truthful or fraudulent. So, that is also the given here. So, this is our data. So, this is a small data set that we have this is the highlighted section here. So, as you can see for observations whether they had prior legal trouble or not yes or no has been indicated appropriately and for the company sizes as you can see whether the company size is small or large that has been specified for every observation here and then the status of these records these observations the specifically financial report whether they have been found to be truthful or fraudulent.

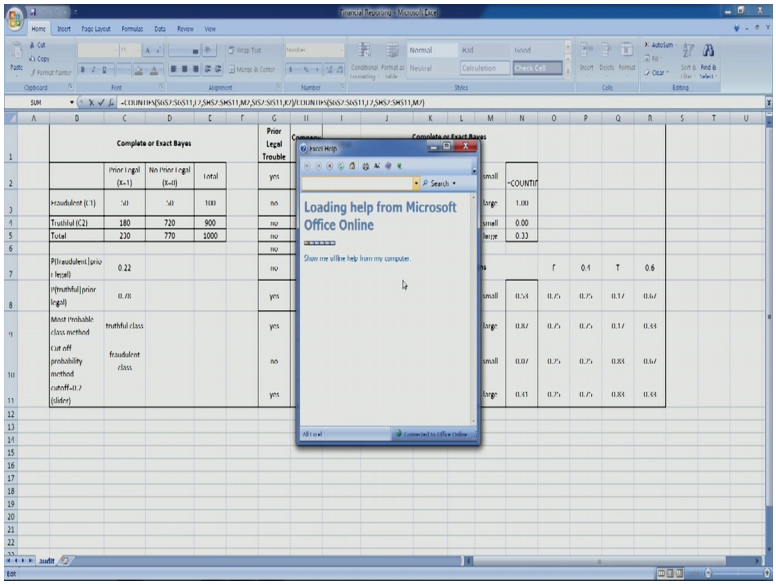
So, that is also specified now if we were to follow complete or exact case modeling. So, in that case how do we go about different calculations and computation? So, let us see. So, let us say we have to compute this probability of a particular record particular financial reports being fraudulent given the prior legal trouble and the company size right. So, in this case because we have just 2 predictors 2 categorical predictors and both these predictors are having 2 classes. So, prior legal trouble we have just 2 classes either yes or no for the company size also we have just 2 classes either small form or large form.

So, given this 2 classes each for these 2 predictors we will have 4 scenarios. So, we will have either we will have these scenarios for the fourth of these scenarios for example, and especially if we are interested in only identifying the fraudulent classes. So, for the fraudulent we have these 4 scenarios when the company when the prior legal trouble is yes and the company size is small. The second one and the prior legal trouble is yes and the company size is large third one when the prior legal trouble is no and the company size is small a fourth one when the prior legal trouble is no and the company size is large.

So, as you can see since we are interested in calculating the probabilities mainly for the fraudulent loss.

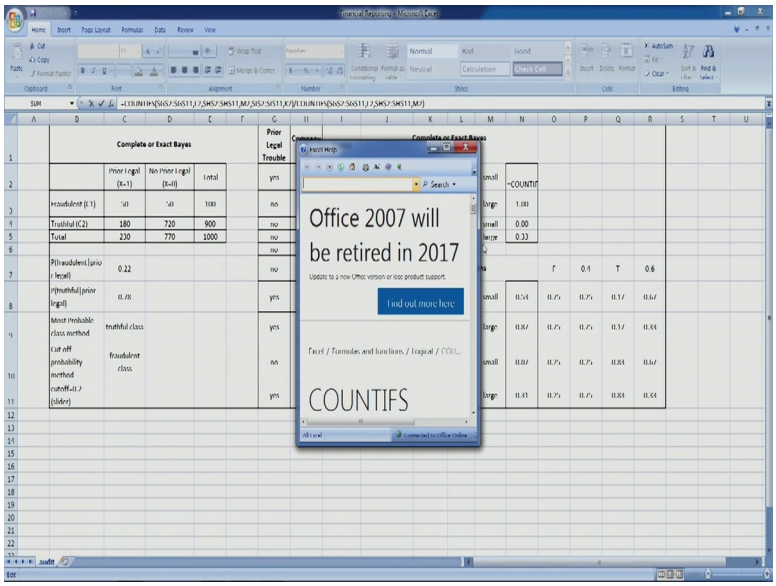
So, this is how we can do it. So, these are the details. So, these details I have further written down in 3 separate columns mainly to perform mainly to be able to use excel functions and to perform computations. So, you can see F the status and then prior legal and then the company size for all these for all 4 probabilities whether that we want to compute. So, I was specified here now to compute the exact based probability this is how we can do it you can see first you can see I am using this countifs function in excel. So, what it does we can look at the details if you are interested. So, you can see a let it load countifs function.

(Refer Slide Time: 06:22)



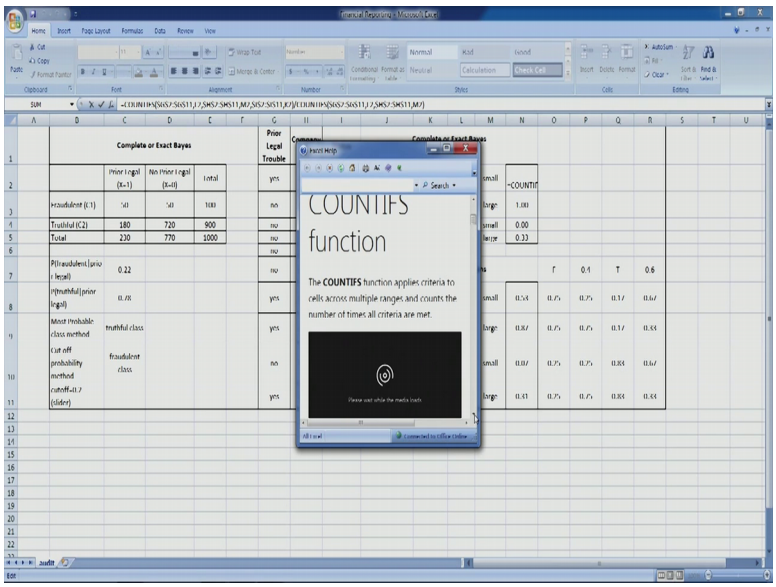
So, this is what we are going to use to perform our comparisons here.

(Refer Slide Time: 06:27)



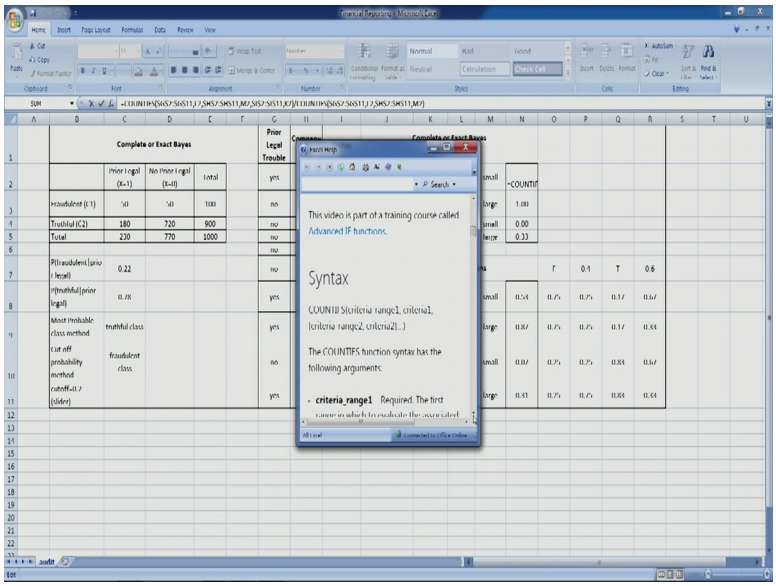
So, function applies criteria to cells across multiple ranges and counts the number of times all criteria are met.

(Refer Slide Time: 06:33)



So, the particular criteria which is specified that is applied across the range and then the count is done.

(Refer Slide Time: 06:50)



So, you have to specify criteria range one and then the criteria. So, you can therefore, in this function you can specify multiple ranges and the associate criteria for all those ranges which is exactly what we want. So, essentially the reason being because we want to compute this probability you can see that the first one here you can see this one prior legal trouble and then followed by the small value that is that is yes.

So, you can see here in this particular colour. So, colour coding is also quite visible here. So, in this first criteria would actually help us identify all the yes observations which we want all the forms which had prior legal trouble. So, we want to identify all those funds. So, there are 1 2 3 and 4 such firms here. So, first criteria and the first range and the associated criteria that is yes we will identify or select or filter those forms then the, if we look at the second one right second one is then the for the company size, the criteria is small here.

So, in this particular column we want to find out all the firms which are having a small size. So, now, in this together these then the further next criteria as you can see is that is whether the firm in the whether the status is fraudulent or truthful. So, you can see the in exchange is that it is only here in the green and within this we would like to apply these criteria that the form is fraudulent. So, these 4 observations they would be identified. Now countifs function will apply all these criterias right and there and then the number would be counted of those observations which satisfy all these criterias.

So, therefore, one has to be yes you know firm has to that prior legal trouble as yes company size is small and status as fraudulent. So, only those observations would be counted here. Then followed by the next one as you can see here the denominator part, new numerator part we have understood. So, all the counts of the funds you know which had prior legal double as yes and company such as small within the fraudulent class. So, that would be counted in the numerator and in the denominator we would have the again 2 criterias. So, first one is that the company had prior legal trouble as yes and then followed by a whether the company size was small.

So, that is the Naive denominator, out of all such firms which had the pretend formation as the prior legal trouble yes and company size being is small we would like to find out the number of firms which all which had status as fraudulent. So, once we execute the particular excel formula then we get the number which is 0.5 in this case if you want to check whether our formula work correctly or not you can do. So, by understanding from this table let us look at the yes and small firms. So, yes I in a small here 1 and then we have 2.

So, there are just 2 cases one yes a small status being truthful and 2 cases yes and smallest is status as being the fraudulent now out of these 2 cases 4 yes and a small right out of these 2 cases based on these 2 predictors information only one of them is fraudulent status I one only one of them is having status of fraudulent. So, therefore, the probability is going to be 1 divided by 2 which is 0.5. So, the other formula is computing this particular exact calculation correctly right in the same fashion again we can compute the exact Bayes probability for the second scenario that is probability for a particular form belonging being filing you know submitting fraudulent financial reports given that they had a prior legal trouble as yes and then companies are also large.

So, the same fashion countifs function using the countifs function available in excel we can compute this particular value as well. So, because this is a small data set again. So, we can find out from the either particular data set itself what is the probability. So, yes and large if you look at the number of observation matching these records as we discussed in the computer exact case you have to find the exact matching records. So, 1 and 2 out of these 2 both are fraudulents, therefore, 2 out of 2.

Therefore, the probability is going to be one the same thing as been computed using the excel formula as well right. So, the formula also the count ifs as you can see a criterion ranges and criteria. So, that has been specified. So, for all the all 3 cells whether the form is no fraud fraudulent or fraudulent reporting or and the prior legal yes and this company size as large. So, these 3 criteria and the numerator and in the denominator of only the 2 criterias that is the predictors information yes and large. So, that would give us the correct exact probability for this particular case.

Similarly for the next case when we want to compute the probability of a firm being fraudulent given the predictor information of prior legal trouble being no and company size being small. So, the same fashion also we can compute as you can see no and is small 1, 2 there are 1 2 and 3 there are 3 exact matches in this particular small data set that we have for the these the situation no and small, but all 3 of them are actually truthful. So, therefore, none of them are fraudulent. So, the probability is going to be 0 divided by 3 therefore, 0 right. Similarly for the last scenario no and large again we can look for the exact matches no and large 1 to large 2 right and lower and large 3. So, in this case, first 2 cases of exact matches no large they were truthful the last one was fraudulent.

Therefore, the exact probability is going to be 1 out of 3 that is 1 divided by 3 0.33 same thing we have computed using excel formulation as well right. So, in this fashion we can compute the complete or exact we can do the complete or exact Bayes calculation and based on that we can perform further steps now if we were to a perform the same steps using a Naive Bayes calculation right. So, how we will do this, the same 4 scenarios that we are interested in identifying fraudulent cases because for accounting firm as we talked about they would be interested in identifying the fraudulent fraudulently reports first because then they can decide about serious auditing or serious scrutiny of those reports right.

So, these 4 scenarios again the same scenarios as for the exact Bayes calculation the same fashion we have written these particular details in 3 separate columns whether for the calculation of the probability. So, in this case you can see how we are trying to compute the Naive Bayes probability. So, in this case we have generated first these conditional probabilities. So, first let us have a look at this conditional probability that have been computed first look at the proportion of records which belong to each of these

classes. So, the proportion of records which belong which are which belong to the fraudulent class these many let us look at the formula here. So, you can see the last column status out of these records which is in the denominator and in the numerator and then we have the criteria which is specified as false. So, out of these how many are actually out of these it costs, how many are actually fraudulent as status.

So, you can we will get the appropriate number which is 0.4 same thing you can compare by just looking at this filter data set being because this is quite a small. So, you can see just 4 out of these 10 observations they are fraudulent. So, therefore, 4 is the probability similarly for truthful class the same fashion we can compute. So, out of this column we can see 6 of the observations and they satisfy these criteria. So, therefore, count if this particular formula that we have written over there it is going to return the value as 0.6.

Now, let us have a look at the other conditional probabilities that we have computed. So, so this is for. So, we have 2 predictors, therefore, and then 2 classes in the outcome variable. So, far first this fraudulent class and these are the 2 for values for 4 values with respect to the predictor this prior legal. So, you can see in the numerator you can see we are trying to identify we are trying to find out the value we are in the as you can see yes where the predictor prior legal trouble is yes right. So, that is the first criteria, that is one that is one criteria prior legal trouble yes and also the record is supposed to be fraudulent.

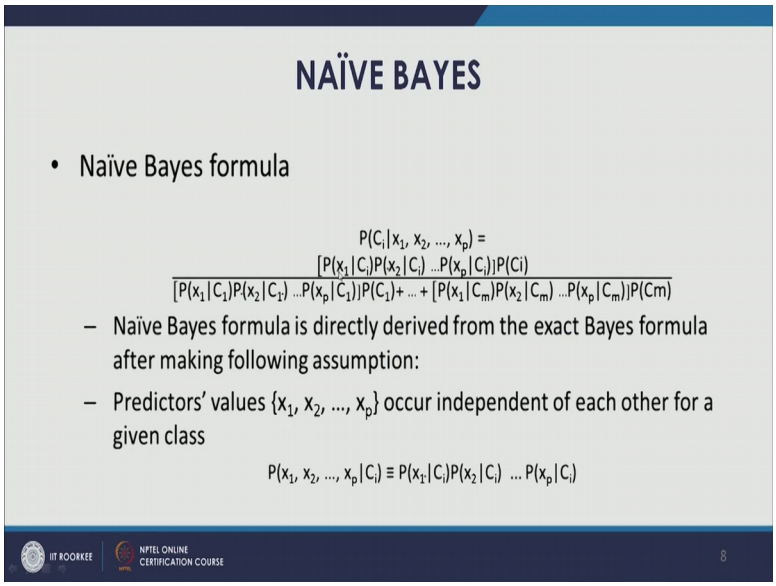
So, these 2 criterias they are in the numerators those counts are to be done and this particular count is then divided by the denominator which is nothing, but the number of records which are fraudulent out of the total records. So, using this we can find out the this particular probability of a record probability of a record being credit for having a prior legal trouble given that it belongs to the fraudulent class, the same thing we can do here in this case this is for the second predictor that is company size.

So, again we can have a look at the formula. So, here also you can see that for these small. So, we have to look at this that the for the for given that this predictor information that the for this particular form this is company sizes small right for this out of the fraudulent cases that we have what is the probability for a firm belonging to that. So, this is comes out to be 0.25. So, after this we can compute the, we can perform the same computations for 3 other scenarios. So, you can see.

So, these values have been computed similarly the same thing has been applied here in the truth for the truthful class. So, here also the same thing is being applied as you can look at the excel formula as well that for the same for the first predictor that is prior legal trouble and the first value being yes now. So, we are trying to count the numbers where in this is particular is predictor value is yes and then the class is truthful out of all the classes which are all out of all the records which are truthful.

So, this value comes out to be 0.17. So, the same thing we have done for other 3 scenarios similarly for the next predictor that is the company size also the same type of computations have been done, now once these values have been computed right.

(Refer Slide Time: 21:54)



So, if we look at the Naive Bayes formula that we have gone through look at the Naive Bayes formula here you can see that and these are the probability values that we are trying to compute here P x 1 given C i P x 2 given C i and the proportion P C i this we have already computed and P x 1 C i and P x 2 C i these values we have already computed right, these 2 values. So, for each class F, these are the values for predictor 1 predictor 2 for class true truthful for predictor 1 and the predictor 2 and then for all the scenarios.

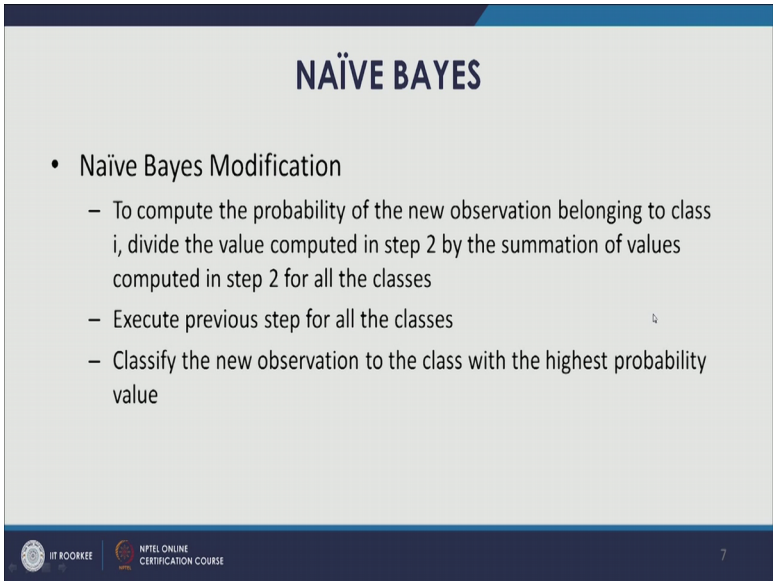
So, we have computed these values right. So, we have been able to compute these values now once these value have been computed we can compute the overall formula for Naive Bayes. So, this is in this particular cell as you can see as you can see from here this is

these 3 cells, you can see these 3 cells are the in the numerator O 8 that is this one this one and then multiplied by P 8 this particular cell and then multiplied by P C 7 in this particular cell.

So, we are trying to compute the probability right of this particular part and then divided by the total. So, that will give us the, this value is 0.53. So, to compute the actual probability as we talked about the first you have to compute the numerator and then you have to divide by the same expression values for all the classes this is what we have done. Similarly for the next scenario as well here also we look at the formula first for the fraudulent class which we are interested in because we want to compute the probability of belonging to fraudulent class you can see.

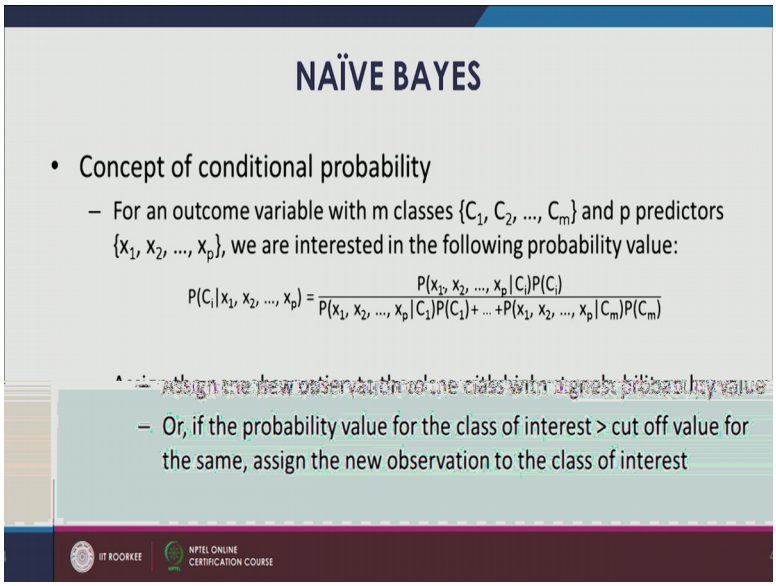
So, these values are O 9 and then multiplied by P 9 and then multiplied by the P 7 value. So, these 3 values and in the denominator we will have this one value plus the other one value which belongs to the truthful class and will get the actual probability as per the Naive Bayes formula similarly for third scenario and similarly for the in the fourth scenario.

(Refer Slide Time: 24:26)



So, exactly following the Naive Bayes formula this one for the Naive Bayes calculation and as we have discussed.

(Refer Slide Time: 24:31)



This one for the exact based calculation now once these computations have been done you can look at now you can do a comparison of these 4 scenarios. So, we look at the complete or exact based values you can find out that this was 0.5 probability of a particular report be belonging to the fraudulent class and given the prior information yes and is small and the same thing you can see the value is 0.5 and here it is 0.53.

The second one the fraudulent class prior legal for years and company size large it is one 4 in exact based calculation 0.87 here in this case the third scenario prior legal trouble no and small form point this is the 0.07 in the exact Bayes this was 0.0 here it is 0.31 in the exact based at this 0.33. So, if we look at the probability where is the actual probability value that we get Naive Bayes from Naive Bayes calculation they are quite close to exact based calculation.

However, we have; however, we do not have to deal with one overwhelming problem that we phase in exact Bayes calculation is that we have to find the exact matches the records exactly matching. So, that we do not have to do because we use the entire data set in Naive Bayes calculation. So, you can see close numbers in both exact Bayes and Naive Bayes calculation and based on these probabilities value then we can decide on you can decide on whether to classify as truthful or fraudulent.

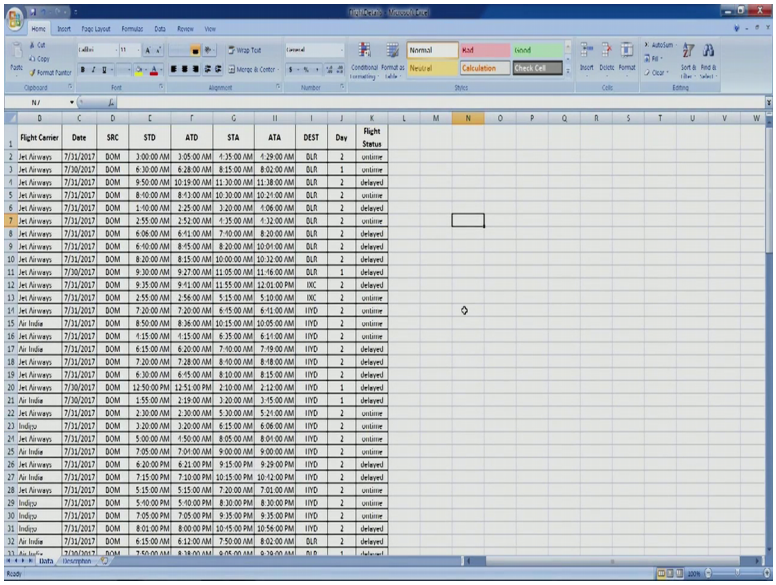
So, these probabilities value if we are not interested if we are not in a rare class scenario then we do not have a special class of interest then in that case we have to compute these

value or values for the truthful class as well for all scenarios and then we can compare and as we had done here in exact Bayes this particular example for the fraudulent as well as truthful he had computed these 2 values and most probable class method can be applied to find out the whether the observation is going to be below assigned as truthful class or fraudulent class cut off probability if we follow the same thing 0.2 right.

Then accordingly in this case also for example, as per Naive Bayes calculation or even for a exact Bayes calculation we follow the cut off tool right using these probability we can assign a new observation to the appropriate class for example, first second and fourth scenario they would be assigned as to belong to the fraudulent class and the last one the third one would be assigned to belong to the truthful class.

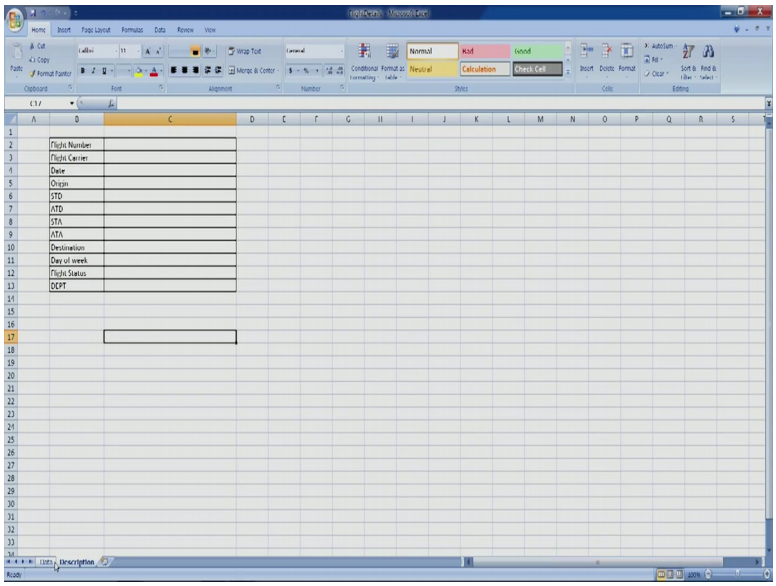
So, with this let us do an as small let us do any small exercise in R. So, what we will do is we will let us get familiar with the data set that we have here.

(Refer Slide Time: 28:11)



So, the data set that we have here is this one. So, this is the data set that we are going to import into R environment and then we will be in doing in exercising R for we will be applying Naive Bayes modeling. So, let us look at the variables that we have these are the variables.

(Refer Slide Time: 28:34)



That we have first one is flight carriers. So, this particular data set is about flights and their carrier the date then source and here then we have schedule time of the departure actual time of departure, then we have scheduled time of arrival, than actual time of arrival then we have destination and then we have day off week.

So, 1 representing here Sunday and 2 representing here in Monday, 2 representing here as Monday, we have just 2 days of information whether the day was Monday or Sunday or Monday then the flight status whether the flight was delayed or on time. So, this has this is based on whether the actual time of dep departure whether that was less than all same as the schedule time of departure if it was less than or same as the schedule time of departure then it is on time if it is more than that then it is delayed.

So, the main problem is the classification task and the predictors as you can see from the data set itself the predictors that we are going to use in the modeling are going to be the categorical predictors and in this particular case the main tasks being the classification task we are trying to predict the status of a flight whether it is going to be depending on the predictors information whether the flight is going to come on arrive on time or it is going to be delayed.

So, we will stop here in this particular lecture and will continue this particular exercise in our in the next one.

Thank you.