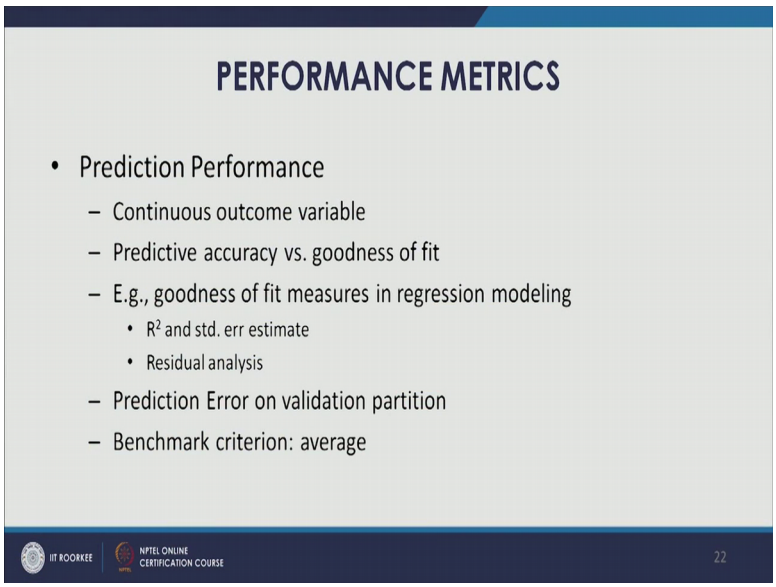
Business Analytics & Data Mining Modeling Using R
Prof. Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture – 21
Performance Matrix-Part VI Prediction Performance

Welcome to the course business analytics and data mining modeling using R. So, in the previous lecture we were discussing performance matrix and we discussed specifically the over sampling scenario, when we have a rare class of interest very few records. So, that was covered we also talked about at the end of that particular lecture, the 2 class scenario where some part some observation are difficult to be classified by our model, and we can we can have another third option cannot say and then that can be a then manually expected by experts and appropriately classified. So, after that we come to our last leg of this particular module also and this particular topic performance matrix. So, will discuss the prediction performance, till now whatever we have been talking about right was applicable to classification performance now we have come to our the last part that is prediction performance.

(Refer Slide Time: 01:18)



So, when we talk about prediction performance essentially it is about. So, we are going to focus on the continuous outcome variable. So, generally we deal with continuous outcome variable. In the classification performance we had categorical outcome variable.

So, most of the things they were dependent on the classes that it have 2 class scenario and extension into m class scenario and all those things, that we have already discussed. Now in prediction performance the focus is now on continuous outcome variable. Now here again we need to understand few things that touch up on few things specifically the classical statistical modelling and data mining modelling, few things we can understand predictive accuracy that is the matrix that we generally use in prediction performance here and data mining modelling, and goodness of fit measures are generally used in statistical modelling.

So, idea main idea behind this difference we have already discussed, but to be specific in this particular topic performance matrix, goodness of fit there in statistical modelling the main idea is to fit the data as closely as possible. We have one sample and we do not do any partitioning the same sample generally primary data is collected in a statistical modelling, and that the hypothesis hypothesis that we have they are tested, and on the same sample we build our model and the same sample that then used is to find out the significance of that model.

So, that is generally done using goodness of fit measures, which typically measure how best that model is fitting the data. While when we talk about the data mining modelling, when we talk about the predictive analytics, we are focusing on how well we can predict the new observation the future data. So, there we focus more on predictive accuracy. So, therefore, in both these settings the scenario is difficult different and the measures that we use to assess the performance of the models are also different.

Now, if we talk about if we like to discuss few examples for classical statistical modelling. So, one would be goodness of fit measures that are generally used in regression modelling in a statistical setting. So, these are 2 measures that we have shown here R square and a standard error estimate. So, generally these 2 measures are used to understand how well the model is fitting the data in a statistical setting.

So, R square as we have talked about this particular metric before as well, in the supplementary lectures that this particular metric measures the in a way captures the variability in the outcome variable. So, how much variability in the outcome variable is actually being explained by the model by the statistical model. So, that in a way in a sense and variability when we talk about, it is about the information spread the spread of

data that is there. So, essentially it boils down to the same thing that how closely the model is fitting the data, how much variability in the outcome variable is being explained by the model or with the help of predictors information.

Standard error estimate that we see in regression modelling that is also in a way telling us how a particular the relationship between a particular predictor and outcome variable, how closely that is following for example, following is being followed by the model in this particular example, the regression model right. So, a standard error estimate would actually indicate that that fitting. Now residual residual analysis is also performed to understand how well the data is how well the model is fitting the data in a statistical setting. We do some analysis we apply some visual visualization techniques as well while we analyse residuals, and we try to find out we try to understand how closely the model is fitting the data and how it is being reflected in the residual series then when we come to a predictive analytics when we come to data mining modelling right.

So, prediction error is one matrix prediction accuracy and prediction error on validation partition. So, generally to performance just like the classification performance that we have been talking about, that is the matrix they are computed on validation partition in the same fashion, even for prediction tasks even to evaluate the prediction performance of the model the matrix, they all computed on validation partitions whether it is prediction accuracy or prediction error, they are computed on validation partition and that is where we compare the performance of different candidate models and then try to select the best model the most useful model using these matrix.

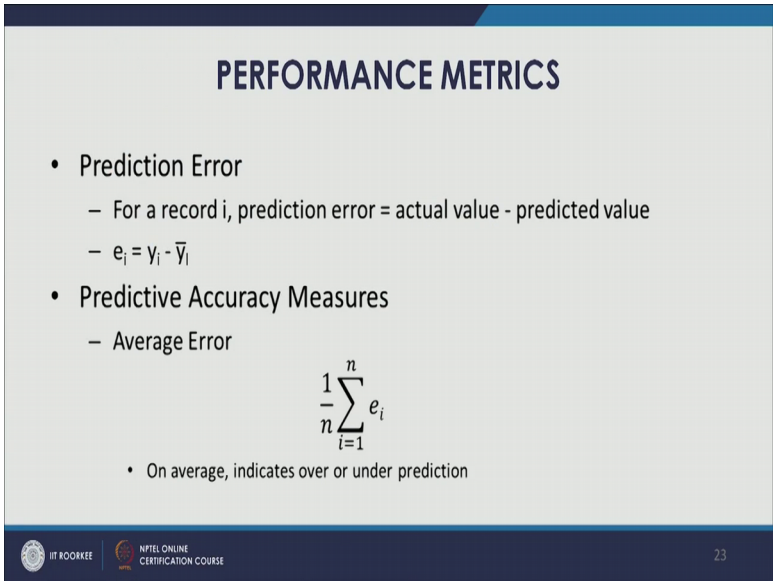
Now, as we talked about in the classification case that naive rule provides us the benchmark the baseline model. Where we talked about be you know if there are m classes and naive rule would be for any new record assign it to the you know most prevalent class. So, that we talked about. So, that becomes the benchmark rule there in the classification.

So, the benchmark rule it does not include it does not incorporate the predators information right when we say that any new record can be assigned to the most prevalent class, then we are not incorporating we are not analyzing the predictor information, and that is used as a benchmark that is used as a baseline model. Similarly in case of prediction right we use average value, we use average value of the outcome variable as

the you know benchmark criterion. So, the average value is actually that that becomes the reference line that becomes the baseline model, and those average value is actually used to compare the performance of the model.

So, all the candidate models that we might build on would be compared to this particular baseline model. Now average value is in a way naive rule equivalent for the prediction task. Now let us discuss few metrics that are applicable in prediction task. So, generally this is the first one prediction error. So, for any record I we can always compute the error.

(Refer Slide Time: 08:34)



So, for any record once we build our model on training partition, and we apply the model on different record that are there in the validation partition or even for that matter in the training partition we are in the test partition. So, for every record the model is going to give us a score right. So, that is going to be the predicted value, and we also have the information on the actual value of that particular record. So, the difference between actual value and predict predicted value that is defined as the prediction error. So, prediction error for a particular record is defined as the actual value minus predicted value. It can be denoted in the mathematical terms as e i as y i minus y i bar. So, in this fashion we can we generally denote this.

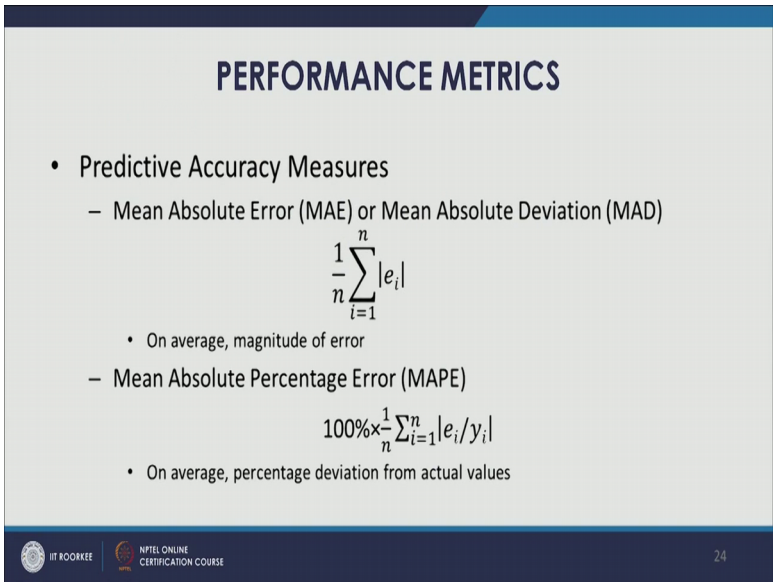
Now, there are some major productive aggressive measures, which are popular which are generally used to assess the performance of a prediction models. So, will discuss them one by one first one is average error. So, if you look at the average error as the name is

suggesting you can look at the formula that is written over there, you can see the residuals for all the records starting from we have and records starting from 1 to n. So, when we say n it could be any partition where we are trying to compute this particular metric average error. So, it could be on training partition could be on validation partition, it could be on test partition, but the performance evaluation that is considered on the validation partition.

So, 1 to n for each observation is starting from 1 to n, we have the that error value that we can compute using the formula that we just discussed before prediction error, and we can summate this this particular these particular values and take an average. So, we can divide it by n this summation and that will give us the average error what does an average error indicate about the performance of the model? So, on average because the this these particular errors they would be at the actual value is bigger than the predicted value, the error is going to be the positive for that record. If the actual value is less than the predicted value error will have a minus sign before it is value, therefore, will have positive and negative errors error values.

So, when we take an average of all the error values for all the observation that are there, on an average level will get either plus or minus sign. So, what does that indicate? So, the average error what does that indicate. So, it will indicate overall under prediction. So, it is a plus sign indicating that on an average level model is over predicting the observations on an average level model is over predicting the observations. If the this average error value comes as minus something then what do we understand from there, that on an average level model is under predicting the value of outcome variable for those observations right. So, under prediction over prediction whether the model is under predicting or over predicting that can be understood from the average error.

(Refer Slide Time: 12:24)

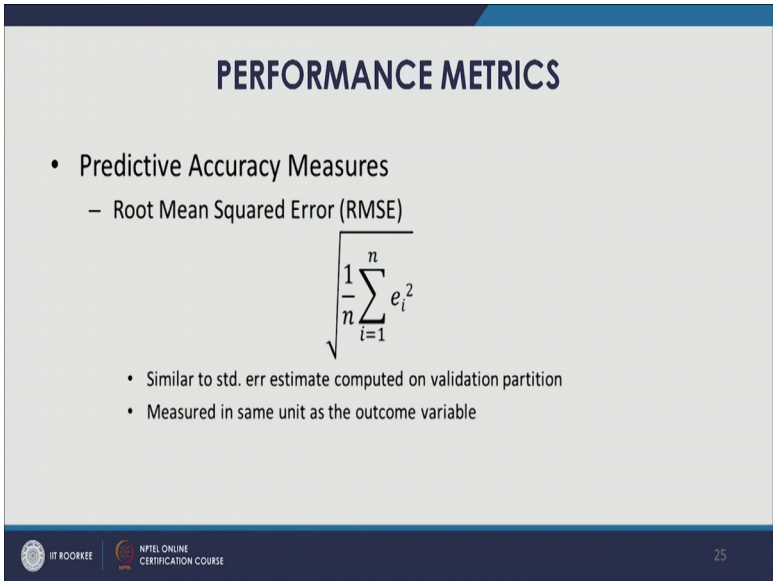


Now, let us discuss next metric. So, next metric is mean absolute error or MAE, sometimes also called mean absolute deviation or MAD you can also see the formula. So, how this is being computed. So, as we talked about that a particular you know a prediction error for a particular record it could be positive or negative. So, because the as the name suggests because we are you know we want to compute absolute error. So, for each error for you know error for each record will we take an absolute value of it, and then again we take the average. So, that is how we get the mean absolute error. So, this particular matrix metric gives us the magnitude of error. So, on an average level we get the magnitude of error, which we are getting from that particular model. So, how much magnitude of error you know in a way for observation this being is coming from the model so that we can get using this particular metric. So, let us move to next matrix. So, next metric is mean absolute percentage error or MAPE.

Now, as the name is suggesting. So, still we are trying to compute the absolute values, but now we are interested in percentage values rather than the actual values. So, to compute the percentage value you would see that the whole formula before the formula we have this 100 percentage, is it is a multiplied by 100 percent. So, that the percentage value can be computed; now we look at the actual expression the error value is being divided by the actual value. So, that we get the difference and then absolute has been taken. So, it could be on either side positive or negative side. So, we take the that difference that quantum that is there, the using error values divided by the actual value

and we take the absolute value of it and then we take the average for including all the observations and then the percentage value. So, what this particular metric tells us about the model that, as you can see in the slide that. On an average level what is the percentage deviation from actual value that is being given by the model. So, what is the percentage. So, how the values or you know the kind of deviation that is the predicted values how what percentage point they are deviating from the actual values for that particular model. So, that thing we can understand using this particular metric.

(Refer Slide Time: 15:18)



Now, let us move to our next metric. So, this next metric is the more important one among the ones we have discussed region will understand. So, this one is root mean squared error also called RMSE. If we look at the formula for computing this particular measure is square root and then we have we are taking the average value of the squared errors, you know we are summating all the squared errors and taking average of it and then taking square root of it. So, you can see when we take a square of a particular error, the sign that we get in a error in the prediction error plus or minus positive or negative sign, that is taken care of now we have a squared values.

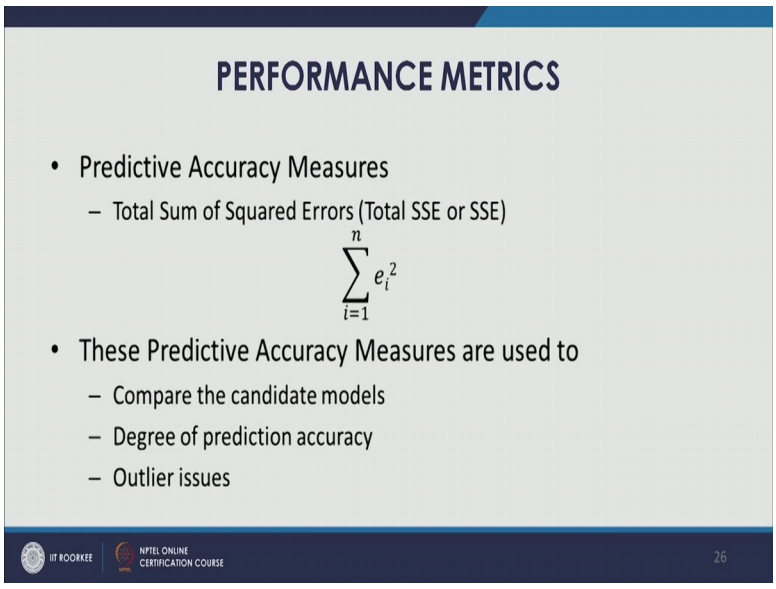
Now all these values are summed up and then be divided by n. So, that is we get the average. So, we get the you know mean is squared error now again we take this square root, once we take this square root again we go back to the same scale. So, this particular scale once we take this square root, it is in the same scale as the scale of prediction error

that is computed on the outcome variable. So, RMSE is very similar to a standard error estimate computed on validation partitions. We talked about the standard error estimate in the context of statistical modelling right we talked about R square and standard error estimate. If we want to have one metric which is quite you know in the predictive analytics this could be used and which is quite close to a standard error of estimate that we get there, this is the RMSE is the matrix. So, this is something that we get for the model.

So, from this we can understand that on an average level, with respect to the outcome variable what is the error that is there. Now this is also another advantage with this root mean squared error is that measured in the same unit as the outcome variable. So, because of these regions RMSE is one of the popular matrix to evaluate the performance of a model. So, whenever we would be comparing performance of different candidate models, RMSE is the one value on which we will rely to evaluate the performance to compare the performance; reason being main that main reason being that same unit as the outcome variable right and quite similar to the standard error estimate that we have in statistical study. So, for that purpose we do use that metric.

Now, the another metric that we might be interested is in total sum of a squared errors sometimes called total SSE or simply SSE as well.

(Refer Slide Time: 18:16)



So, how it is computed simply all the errors square values of those errors, and then summation for all the observations. So, that computes us for the total SSE. So, this what does this particular matric indicate? So, this particular metric indicates us the gives us the total you know that overall sense of error that is being given by the model. So, if we are comparing 2 models, then we can look at the SSE value how much how what was the SSE value for model 1 model 2 model 3 and onwards and the this particular value will also give us the sense that which model is you know performing better. Though the scale of these errors would be different, but comparison is still feasible.

So, this is to give the overall sense of the error for a particular model. Now these particular matrix measures that we have talked about for prediction tasks. So, where we can use them? So, we can use them to compare the candidate models as we in discussing we can also assess the degree of prediction accuracy that is there. Now what is the problem with these matrix is there a are the matrix can always be use can there be a few issues outlier related issues could also be there. For example, all the matrix all the formulas that we talked about they generally consider all the observation, and when we they consider all the observation they include all the error values all the prediction error values, a irrespective of whether it is lying within the major majority of the values or whether it is it is an outlier value. So, that might complicate the assessment of models.

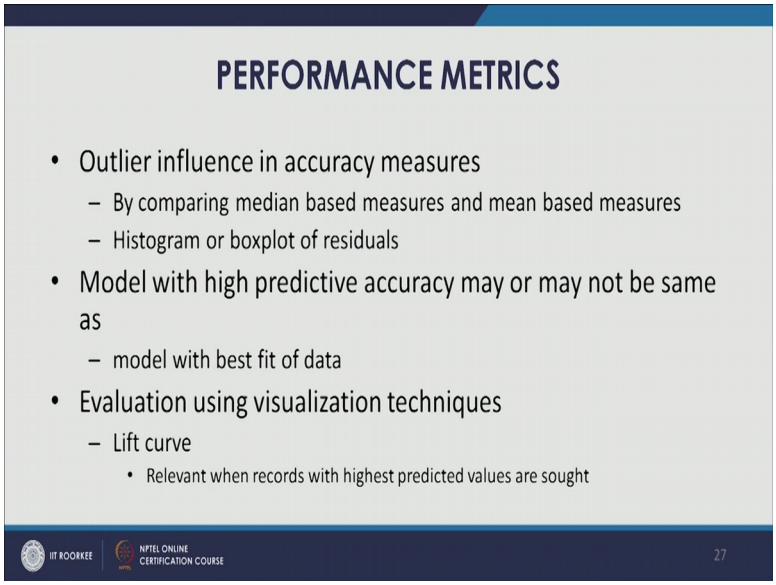
So, how do we overcome such a situation, how we how do we overcome outlier influence in our measures that we have discussed. So, we can compare some we can use some of the median based measures and we can compare them with the mean based measures. So, from that comparison just like what we do in when we check for normal distribution and whether it is right skewed, or left skewed we look at the median and mean values. If mean value is higher than medium then probably it is right skewed and if the reverse is true then it is left skewed, in the same fashion we can compare the median based measures and the mean based measure and from there we can get a sense whether outliers the kind of influence that outlier have.

Another way would be to apply visualization techniques for example, we can use histogram or box plot of residuals we can plot we can generate histogram and box plot of residuals and from there we can you know the same kind of thing we can observe there what we do for normal distribution when we check for skewness. The similar kind of observation we can make and then check for the outlier influence.

Now another point that we need to understand at this point is that when we build a model in a statistical setting, and when we build a model in a data mining setting and in data mining setting we are going for the higher predictive accuracy and when we talk about statistical setting we are going for the best fit of data. So, the model that we get in these 2 different settings they may or may not be same. So, the model that we get when we build a model and data mining setting and using the matrix using following high predictive accuracy, it might be different from the model that we get when we you know build a model in a statistical setting, while we are looking for best fit of data.

Now, can there be a visualization techniques which could be used to evaluate the performance of you know prediction models. So, lift curve is one which can be used.

(Refer Slide Time: 22:35)



Now, lift curve that is the exposure that we have till now, we have been using lift curve quite often. So, at it is going to be relevant only when records with highest predicted values are sort. So, as we have seen before that channel lift curve we compute some values, and then we take accumulated numbers and then those accumulated values are then plotted against the number of cases there, and then we try to assess how much lift we get in comparison to the baseline model.

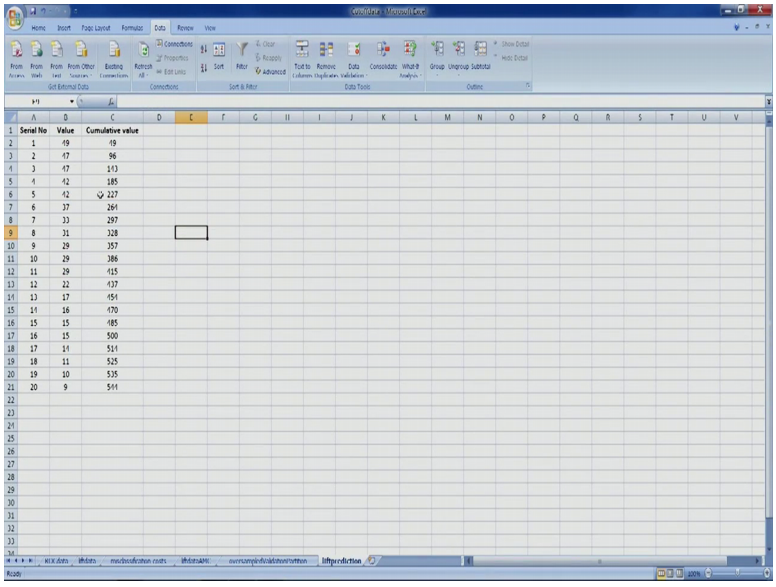
So, in situation in prediction tasks if we have a situation wherein for example, sedan card data set that we have been using. If we take an example of sedan car and there is a company and they have multiple channels to sell their sedan cars right. So, different

channels you know might give them different kinds of revenues right. So, and especially the data set if it is for example, if it is for used cars right. So, therefore, it becomes difficult to assess the price, new cars this is fixed from the manufacturers side so, but when we talk about used cars the prices might vary and therefore, different channels how different channels can be used by the firm to sell those cars.

So, lift curve can be useful in that sense for example, a particular firm might have few of it is own channels and few a few it might also use some of the channels, operated by third party. So, the firm might look to retain you know look to sells the highly priced cars through it is own channel. So, that it can make more revenues and therefore, more profit using it is own channel and some of the low value used cars they might go for with the third party channels.

So, that can be done using lift curve. So, will do an exercise in R to see how this is how this can be done.

(Refer Slide Time: 24:54)

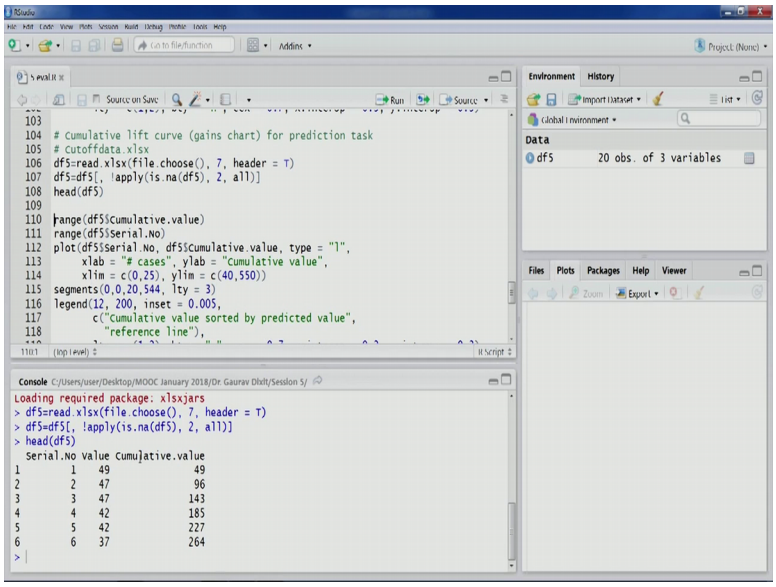


So, this is this small data set that we have you can see we have already this these a column called value, serial number is there then value is there this value of car can be considered for the cars sedan cars premium cars. So, they are ranging from 9 lakhs to 50 lakhs, 8 lakhs to 50 lakhs. So, random functions were used to generate these values and then these values worth then later on sorted in the decreasing order. So, as you can see

once they were sorted, then the cumulative scores were a also computed as we have been doing before as well. So, you can see the cumulative scores have been computed.

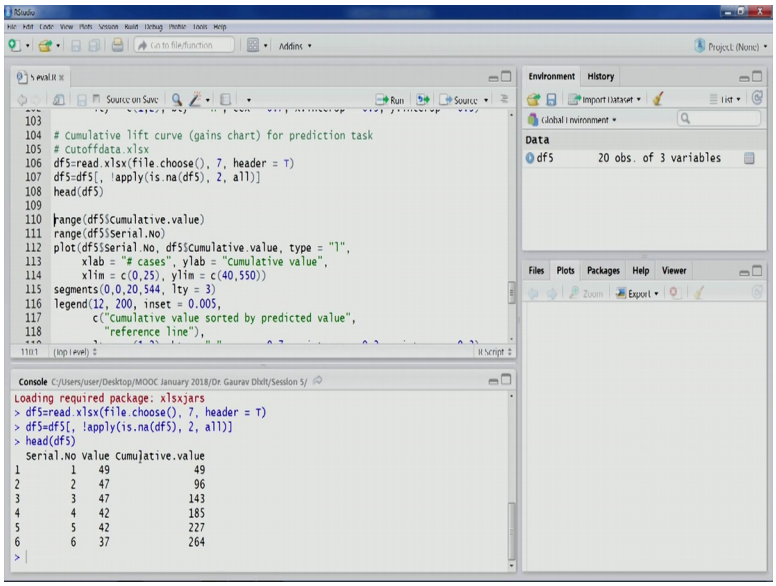
So, this is how you can generate the data, now what we are interested in we are interested in the for example, higher value cars. So, which ones are the which first 10 cases are you know the having the highest value right. So, we would like to identify those high value cases right.

(Refer Slide Time: 26:01)



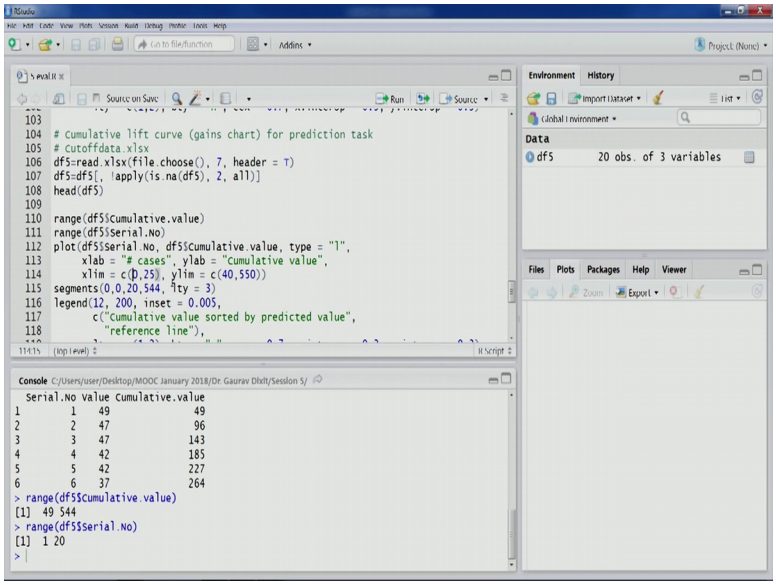
So, let us open R studio let us load this library x l s x, now the data set that we have just seen it is in this particular file cutoff data, let us import this particular data set. Let us remove the na columns and first six observations this we have already seen these observations cumulative value. So, we would like to plot these cumulative value with respect to the serial numbers that are there.

(Refer Slide Time: 26:32)



So, let us look at the range that is there. So, the range is between 49 to 544. So, this is the cumulative value range that we have serial numbers I think we have just 20 records.

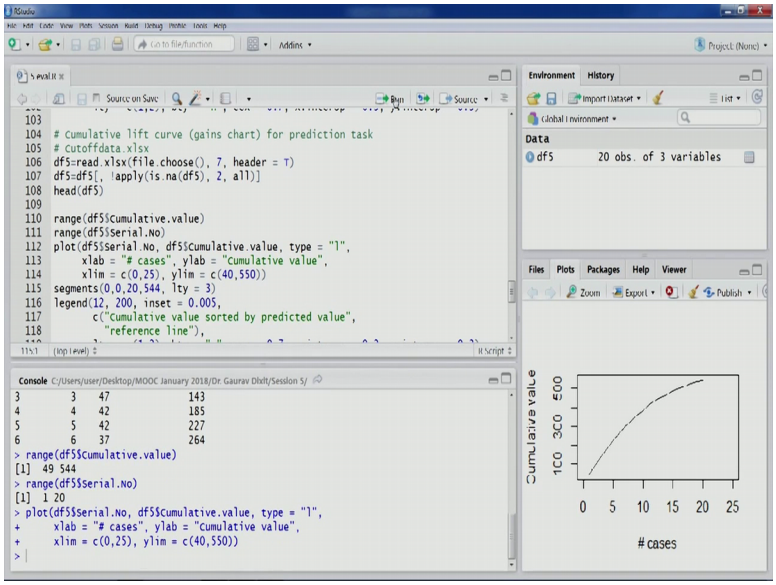
(Refer Slide Time: 26:47)



So, range is going to be 1 to 20. So, you can see the plot the way we have been doing before or as well the first variable first argument is going to be plotted on x axis that is serial number, then the y argument second argument cumulative value that is going to be plotted on y and the type of the curve is lined and labelling for x axis and y axis cumulative value for y axis is given, limits have been appropriately specified as you can

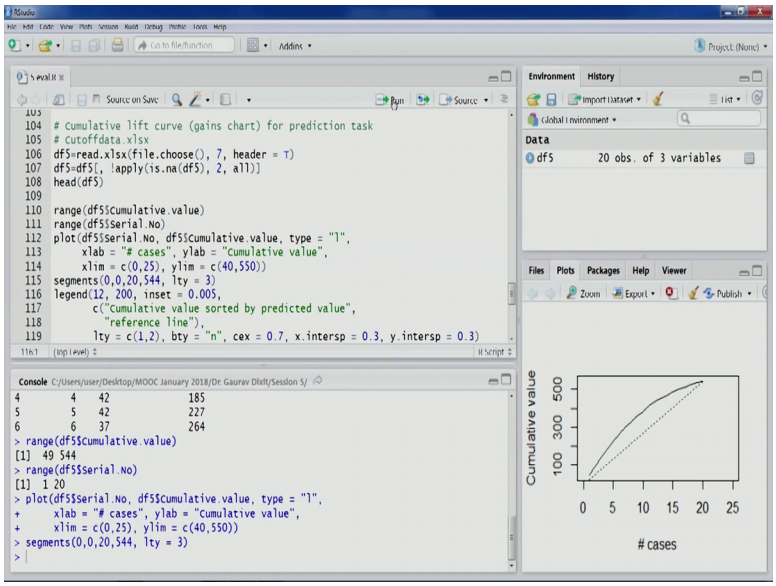
see range we have already computed at 0 to 25 and then we have this 40 to 550 which is covering the entire range that we saw.

(Refer Slide Time: 27:38)



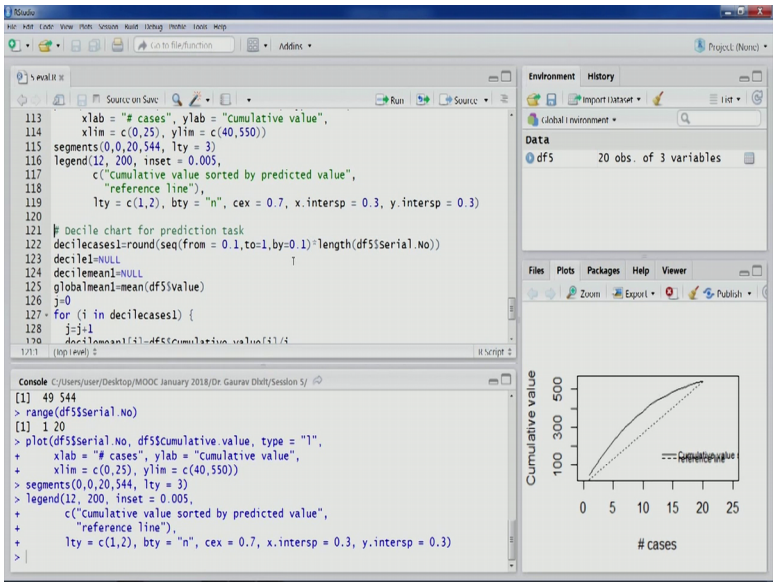
So, let us execute this line of code and so, this is the plot that we get. So, as you can see this is smooth plot lift curve has been created now let us also draw the reference line. So, reference line would be connecting the initial point with the last point that we have let us draw this.

(Refer Slide Time: 27:59)



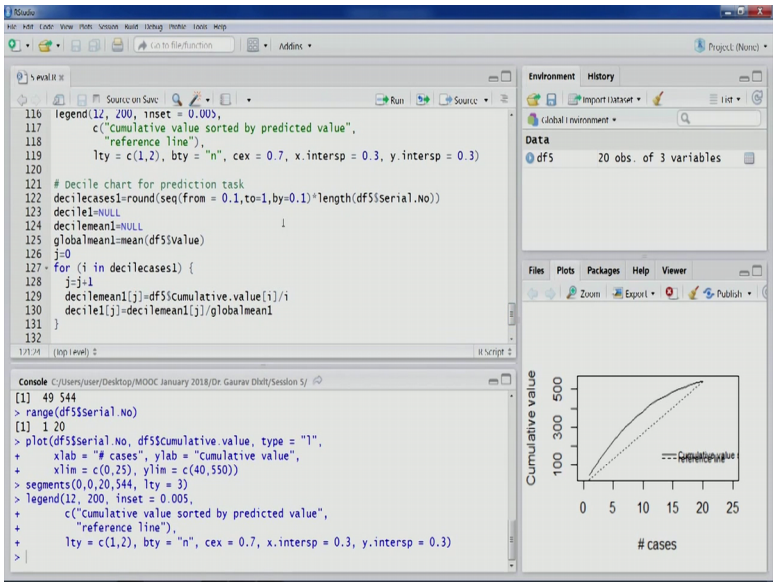
So, this is the reference line. So, this lift line that we have. So, there is good enough separation between the lift curve and the reference line. So, therefore, model is useful and giving us some effectiveness in terms of identifying those high value cars right. So, high value used cars sedan cars it is the model is giving us some usefulness and we can have those cars and sell those cars using our own channels if right and the other cars the low value cars we can probably post them through third party channels.

(Refer Slide Time: 28:47)



So, let us also create the legend you would see the cumulative value is started by predicted value from this lift curve and in the reference line. Now the information that we have just defected using lift curve can also be done using the decile chart like we had done before for classification tasks. So, let us create a decile chart for prediction task. So, in this case also because deci we would like to have 10 deciles and each decile representing 10 percent of the cases, and 20 percent, 30 percent and so on.

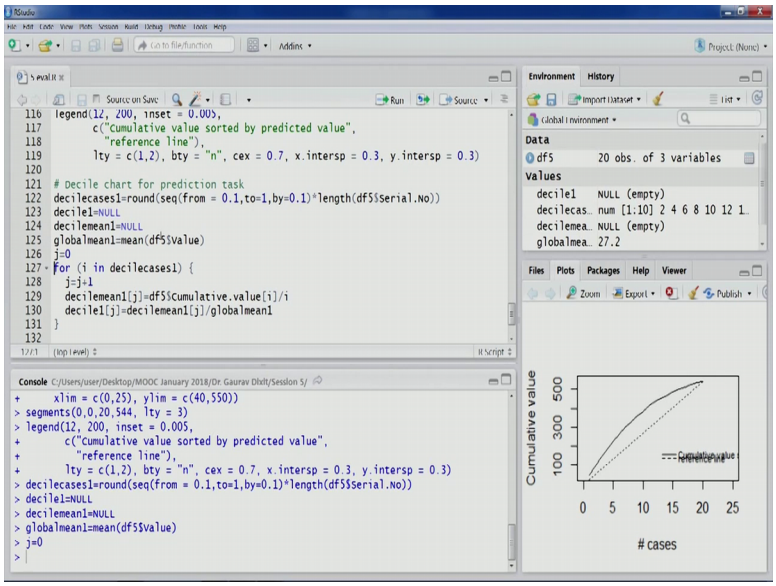
(Refer Slide Time: 29:09)



So, we would like to create this variable which will have the information on the f d decile the number of observation that would be there. So, you can see sequence and the length is of this particular variable which will give us the total number of observation and that that would be appropriately distributed for each decile you would see that 2 4 6 8 10 12 in this fashion will have the distribution.

Now, few other variables decile that is where we would be computing the decile values, that would be plotted using word chart later on decile mean. So, this we want to compute. So, let us initialize this value as null now global mean in this case would be you would remember how we computed the global mean for the classification tasks, now here we just need to compute the average of this particular column and will get the global mean. So, global mean comes out to be 27.2 lakhs now let us initialize these counters.

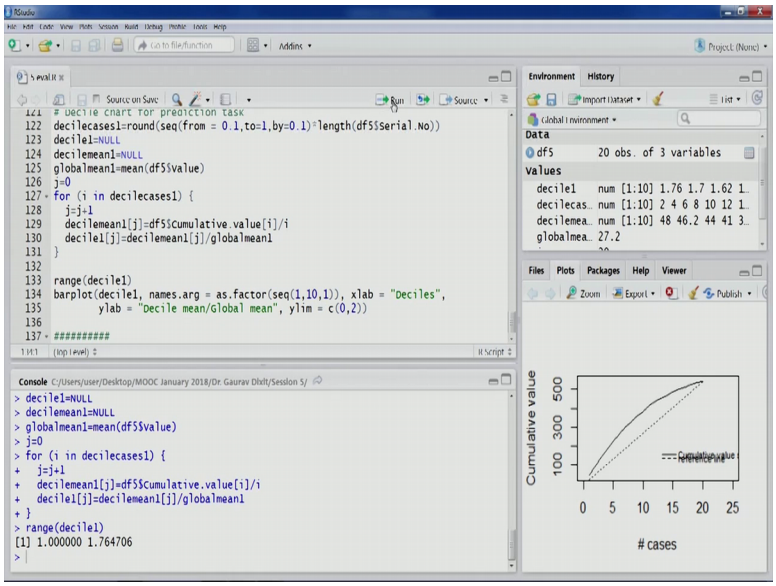
(Refer Slide Time: 30:17)



Now, we are going to run this particular for loop starting in decile cases where all the decile 10 deciles are there. So, for each deciles this loop is going to be run and we would be computing and every loop decile mean, which is right for every decile this is actually nothing, but the cumulative value that we have up to that decile and divided by this particular number of that decile, right.

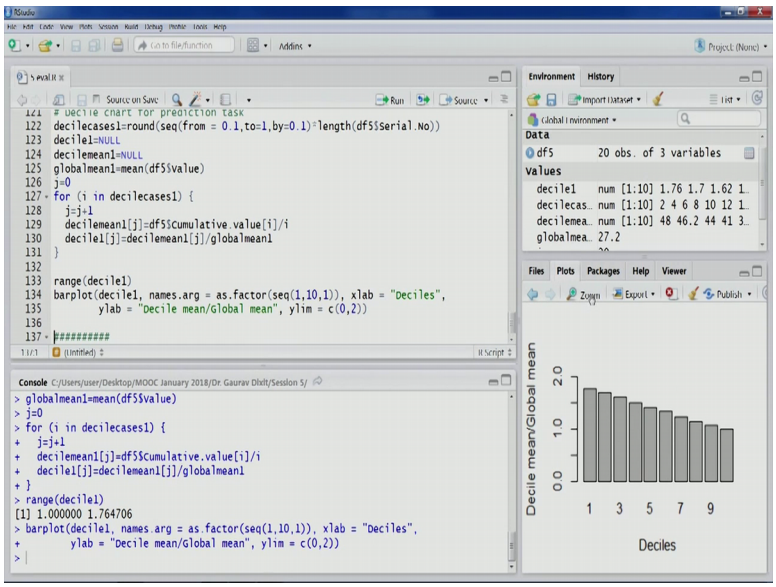
So, that will give us the decile mean and then we can divide the decile mean by global mean and then that will be used as our decile value. So, let us execute this particular loop and will have all the decile values here, you can see all the decile value starting from 1.76 in the decreasing order they are going to be there.

(Refer Slide Time: 31:10)



So, let us look at the range the range is from 1 to 1.76 now we are going to generate this bar plot, which would be creating this decile chart. So, decile one these values decile values that we have just computed while limit is representing the range 0 to 2 is going to cover that deciles level of the x axis, then the level of y axis decile mean divided by global mean and the arguments appropriately specified for each decile let us compute.

(Refer Slide Time: 31:39)



So, this is the plot that we generate this is the decile chart that we have. So, as you can see decile one representing 10 percent of the cases. So, the value as we have competed it

is 1.6; 1.76. So, it will give us the liftoff 1.76 in comparison to the baseline model in comparison to the random selection scenario. So, these probably these 10 percent cases see we would be interested in selling these used cars these premium sedan cars using our own channels, because we might generate more revenue more sales through this. Similarly for decile 2 also we get a good enough lift value right is also more than 1.5 it is actually 1.7.

So, in this fashion we can find out the optimal you know we can take the appropriate decisions till the appropriate lift value. So, lift value of more than one would be useful for us. So, in this fashion we can find out the number of cars that we would like to sell through our own channels. So, thank you will stop here. So, that concludes this particular module and in the next lecture will start the next module that is on supervise learning methods and the very first technique that we are going to start on is the multiple linear regression.

Thank you.