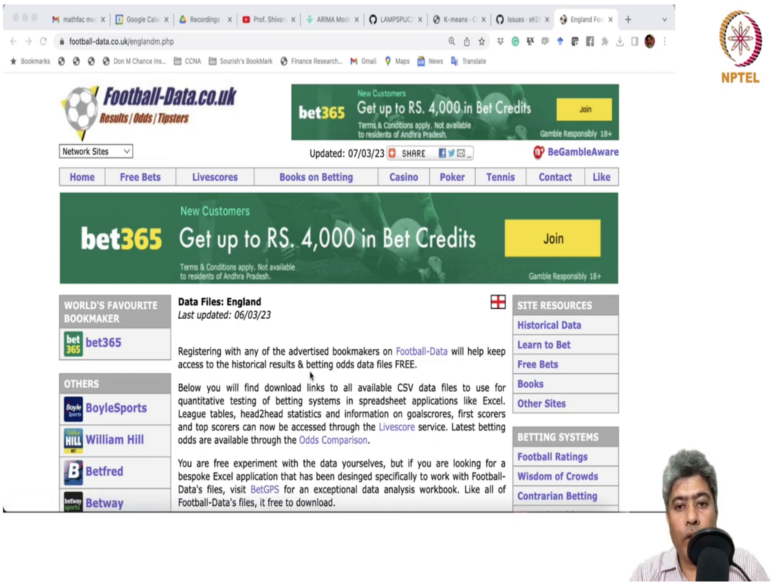
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Mathematical Institute, Chennai

Lecture - 53
Hands on with R: Implement Tree Regression and Random Forest with EPL football Data

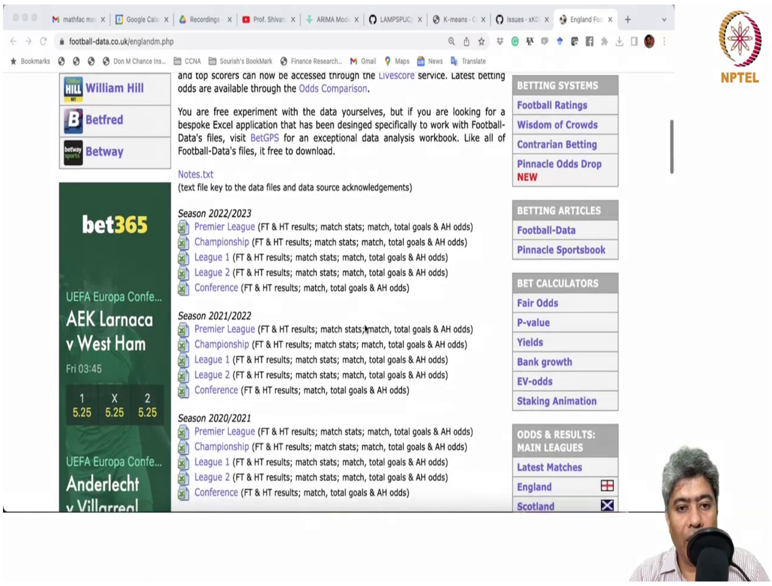
Hello all. In this video we are going to Implement the Regression tree and Random Forest using R. We will use real life English Premier League dataset and we will see; we will also check how good these models are in out of the sample.

(Refer Slide Time: 00:43)

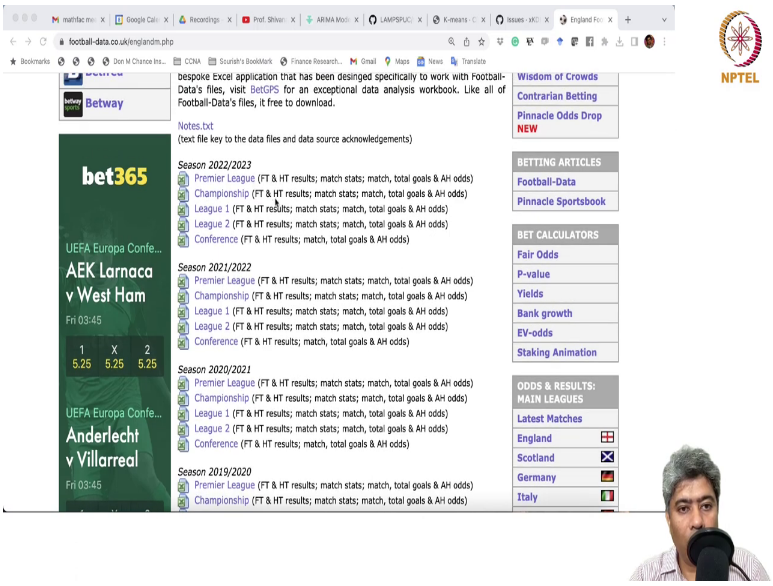


So, I am going to take call these, I am going to you know directly call these datasets.

(Refer Slide Time: 00:49)



(Refer Slide Time: 00:54)

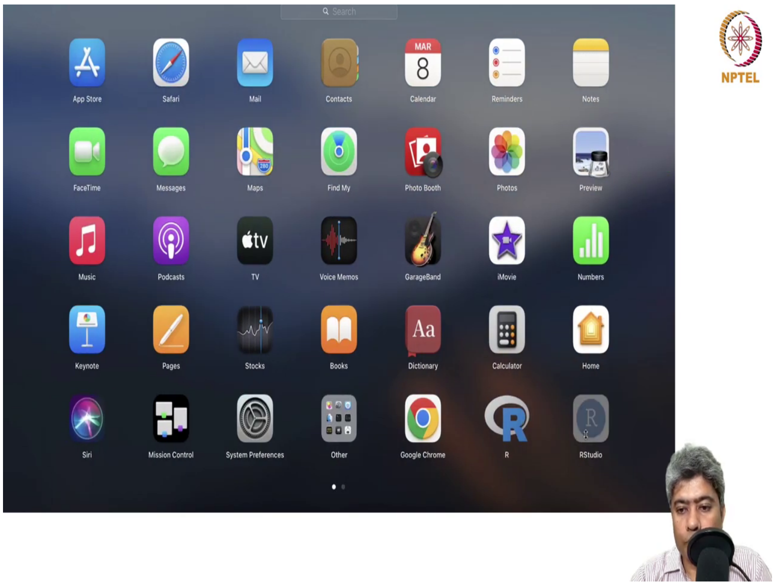


So, there is 22, 23. I am not sure if all the data available for the last league. We can check it out. Let me just see. So, what we I will do? Let me first open R.

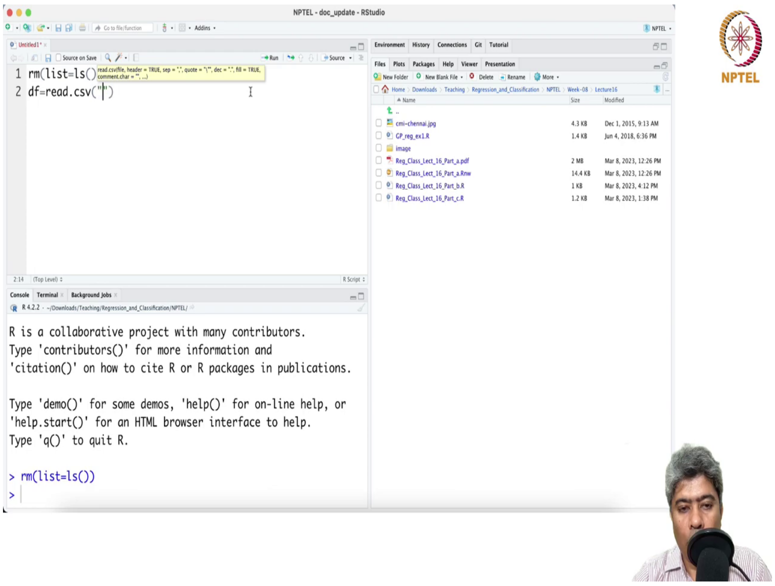
(Refer Slide Time: 01:12)

Open R.

(Refer Slide Time: 01:17)

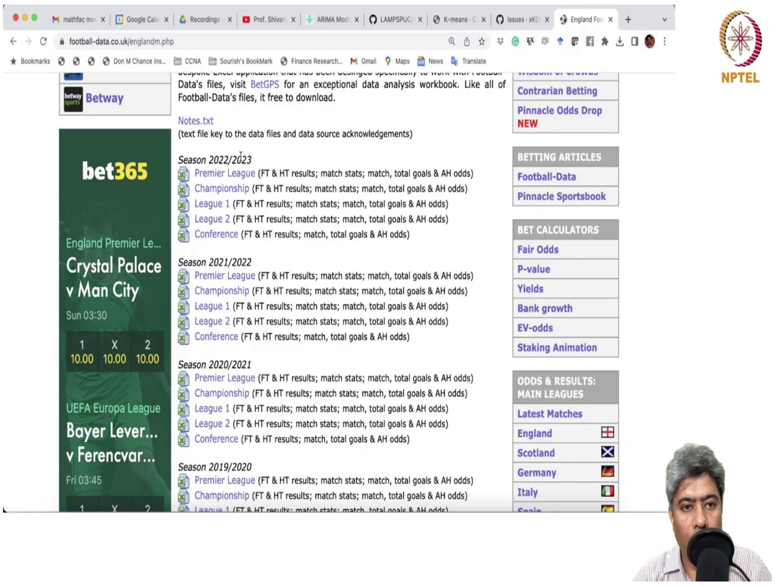


(Refer Slide Time: 01:20)



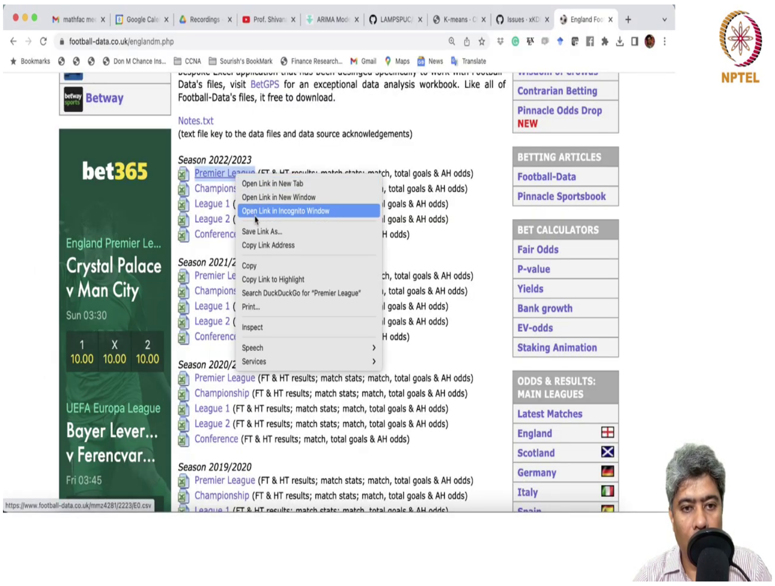
So, first what I will do df say equal to, let me first write rm, list equal to ls. So, it always a good practice to have that start with a clean environment and then data frame read dot csv.

(Refer Slide Time: 01:54)



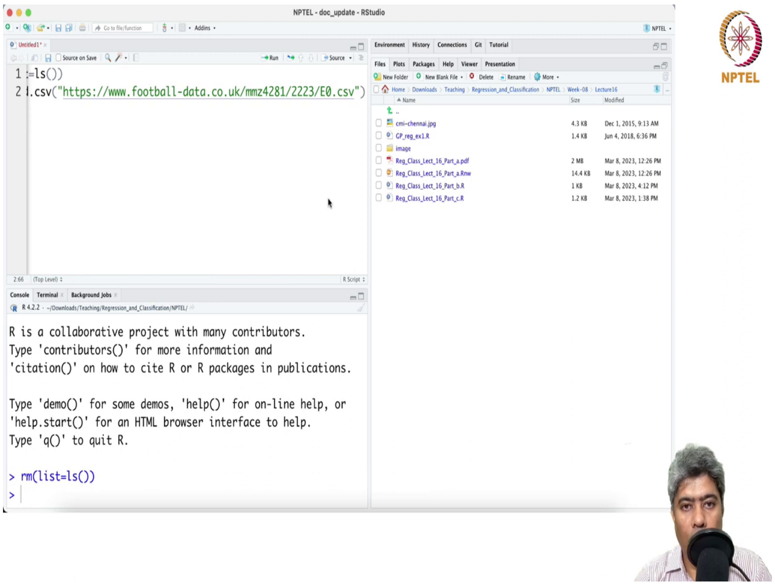
First what I will do.

(Refer Slide Time: 01:56)

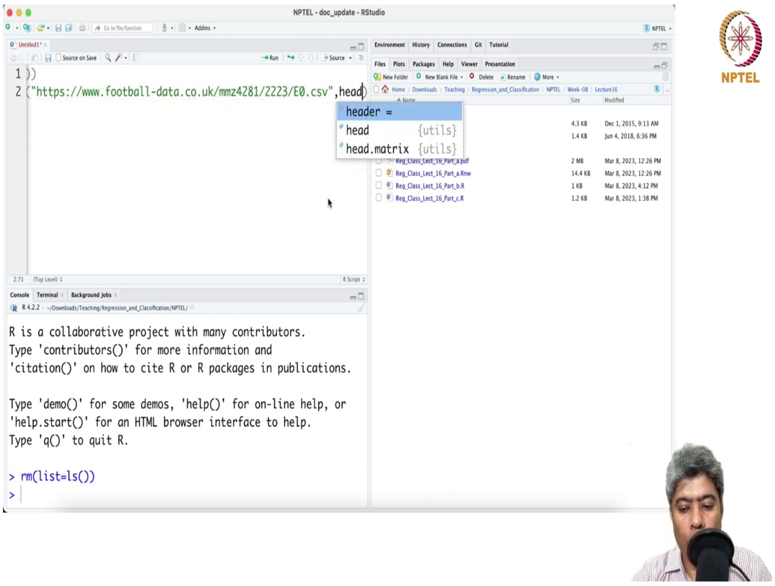


I will just take the copy the link address and I will put it here and head equal to true.

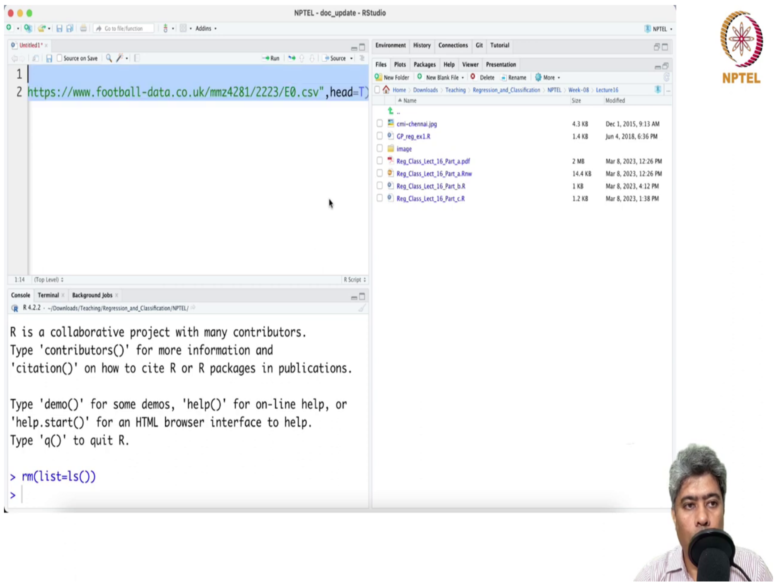
(Refer Slide Time: 02:02)



(Refer Slide Time: 02:05)

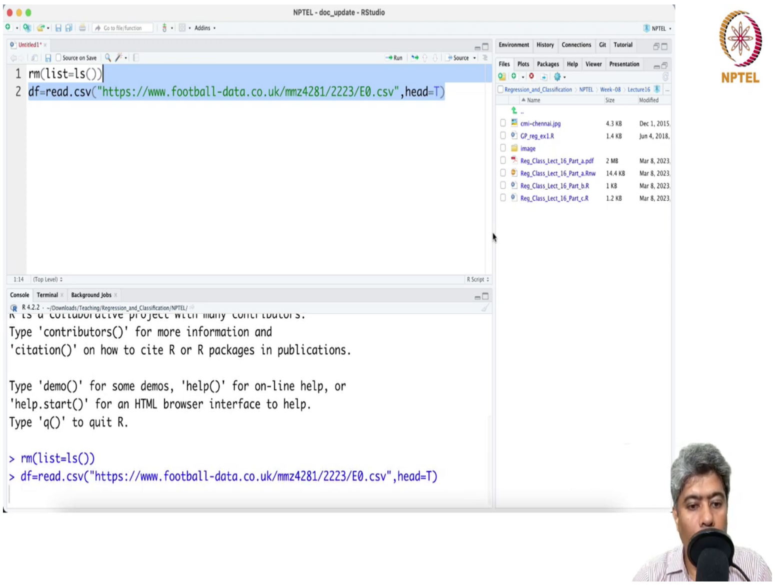


(Refer Slide Time: 02:08)

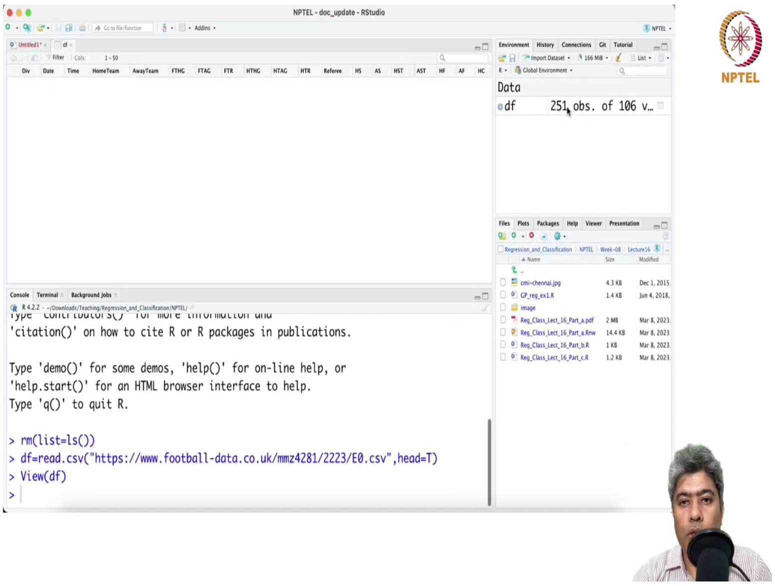


So, yeah let me run that.

(Refer Slide Time: 02:16)

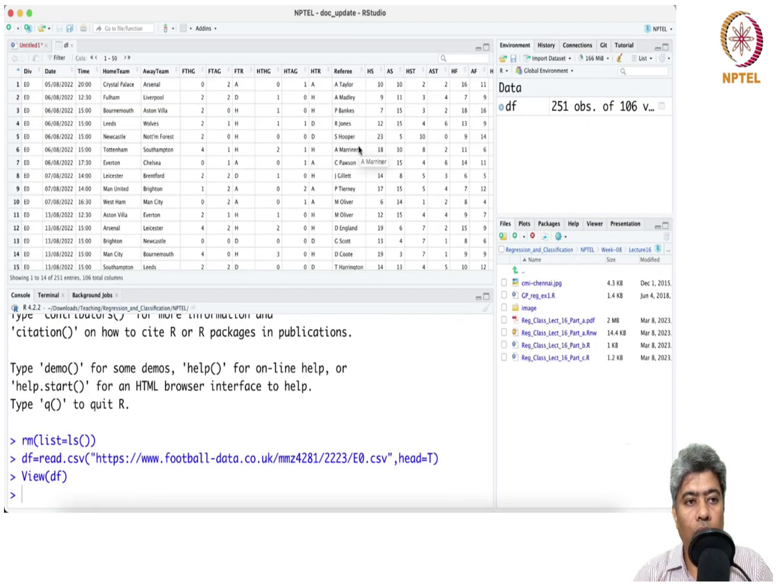


(Refer Slide Time: 02:20)

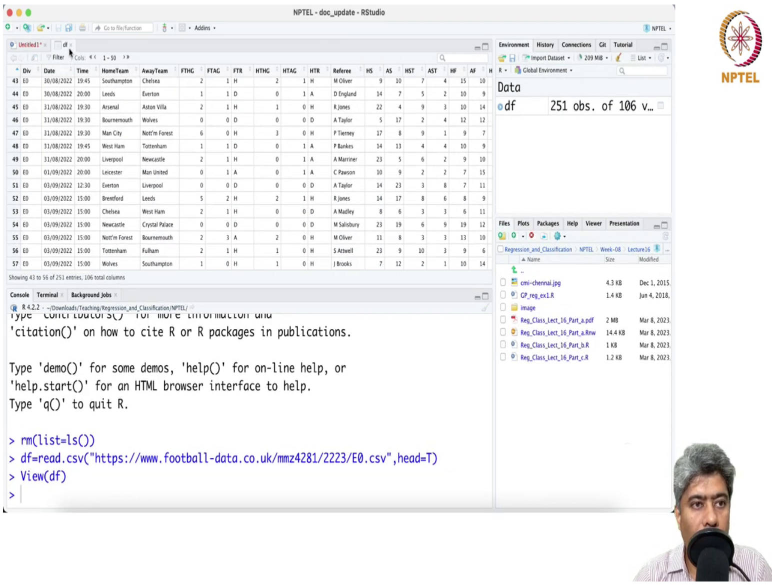


So, here is. So, there are 251 observation.

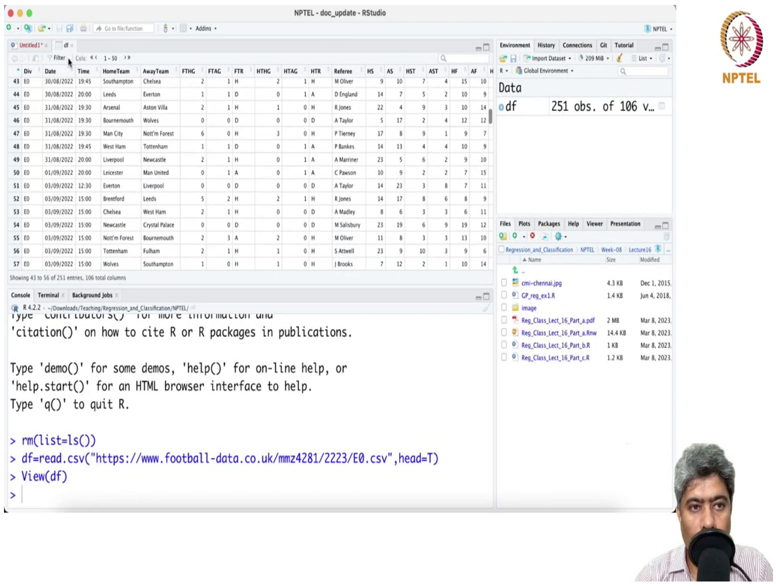
(Refer Slide Time: 02:21)



(Refer Slide Time: 02:23)

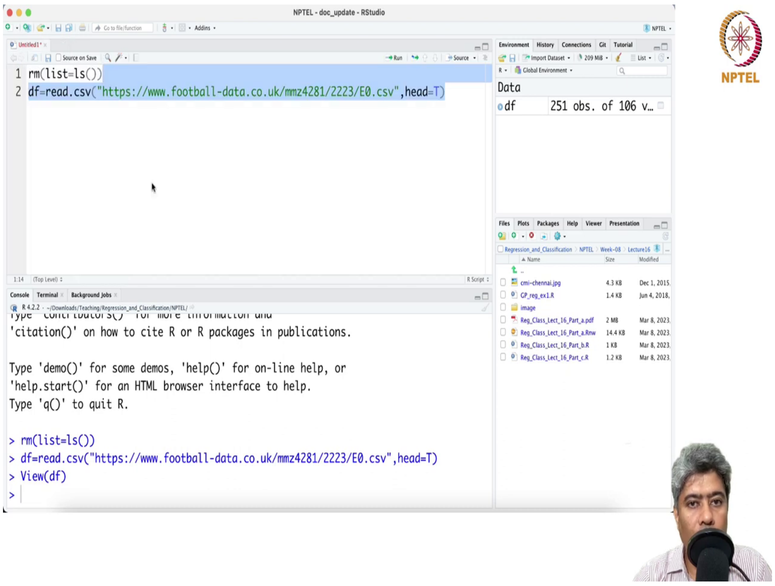


(Refer Slide Time: 02:23)



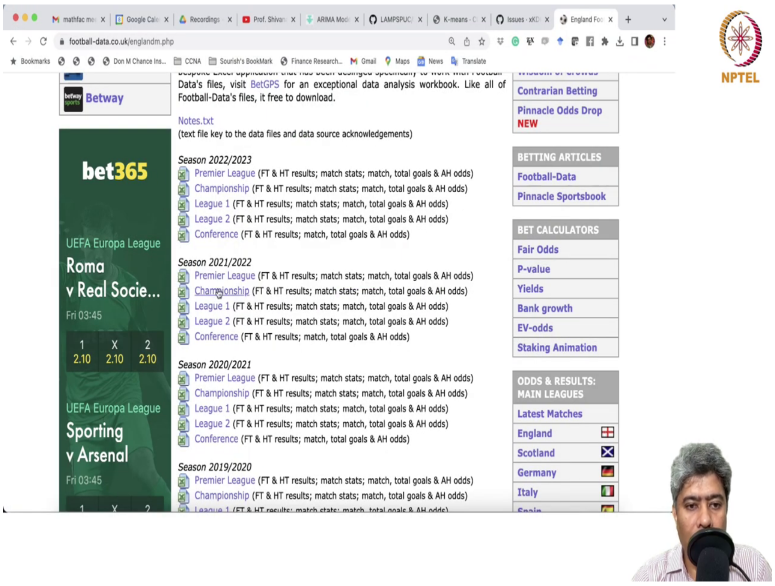
That means English Premier League is all still running.

(Refer Slide Time: 02:25)



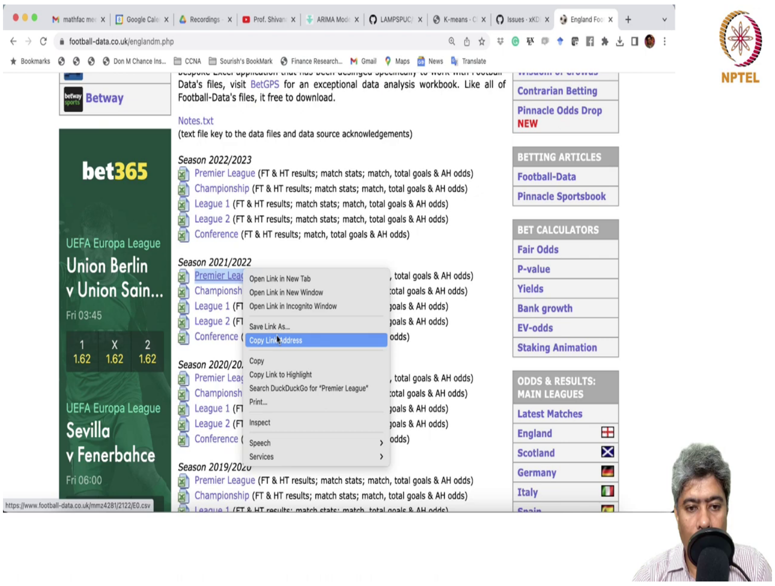
So, they are not all observations are available, ok. So, not all results are available. So, what I will do?

(Refer Slide Time: 02:36)



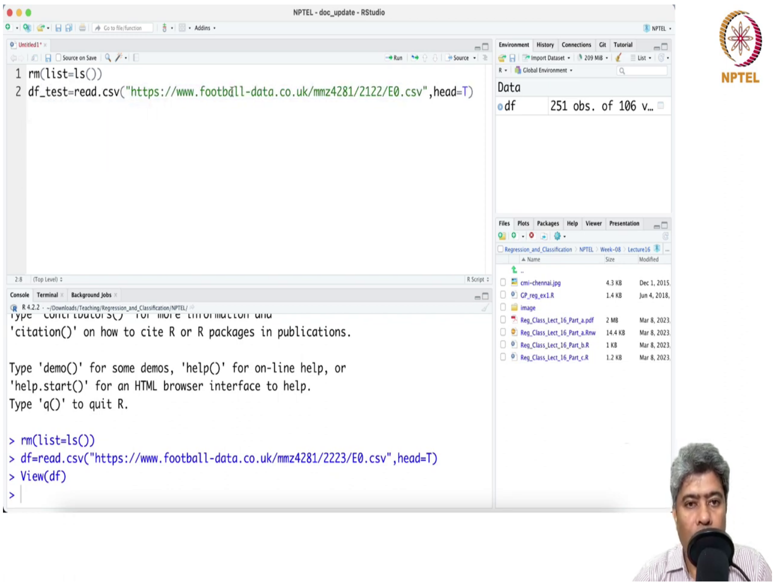
Let us instead of that, let us take 21, 22 data. So, we will take 20, 21 seasons data and train the model and we will use 21, 22 data to test the model.

(Refer Slide Time: 02:55)



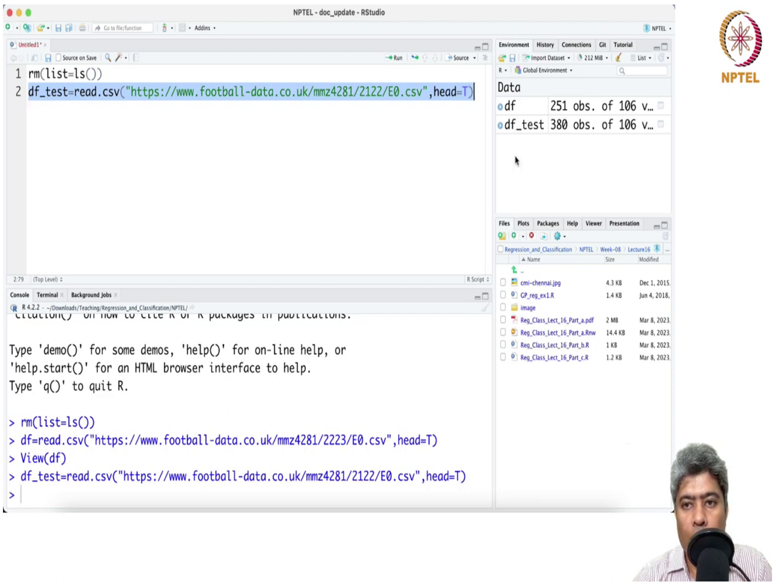
So, let me link address. I will just do that.

(Refer Slide Time: 03:06)



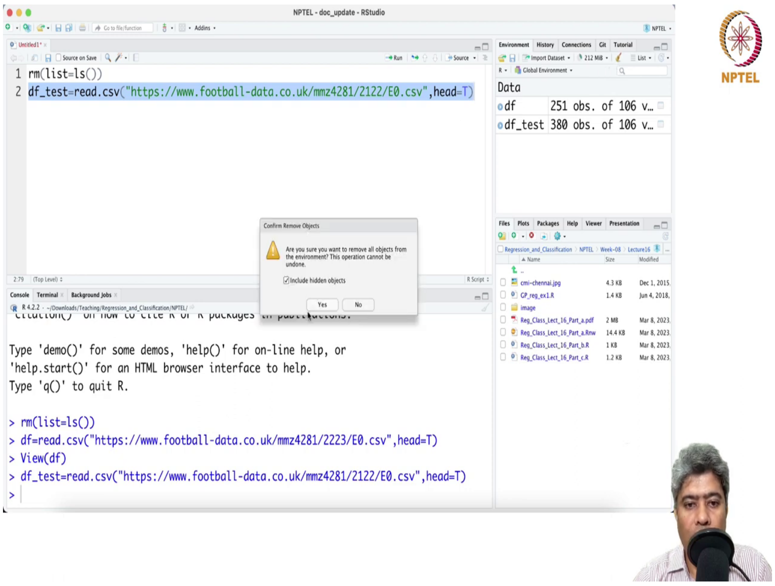
So, this is our test data set.

(Refer Slide Time: 03:14)



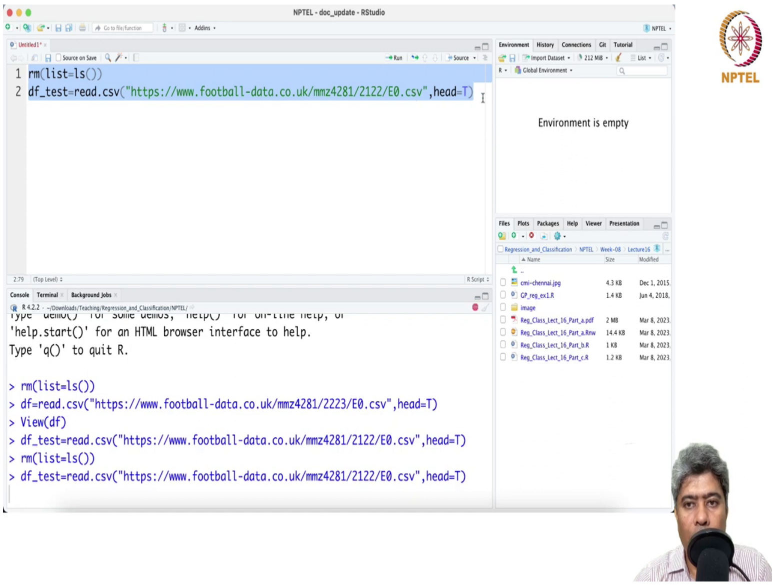
So, you can see 21, 22 season, we will use it as a test data set, ok. So, you can see that 380 observations are all there.

(Refer Slide Time: 03:19)



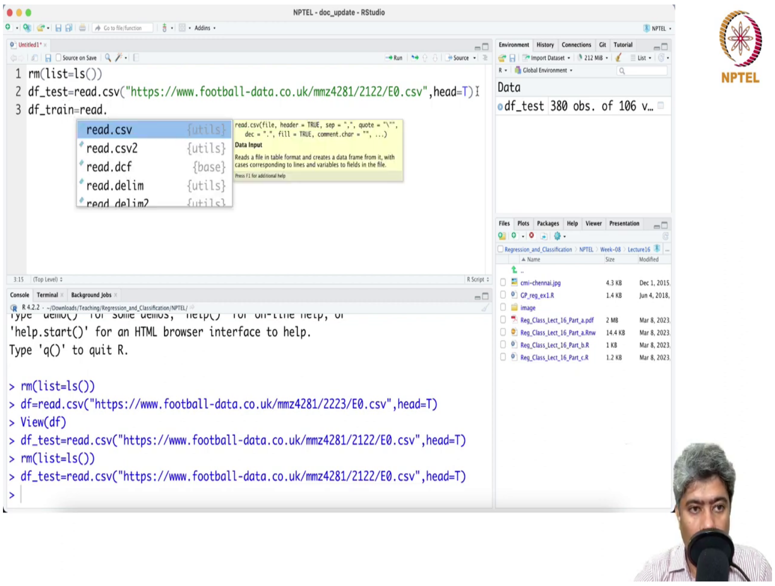
So, you can just clean this overall.

(Refer Slide Time: 03:23)



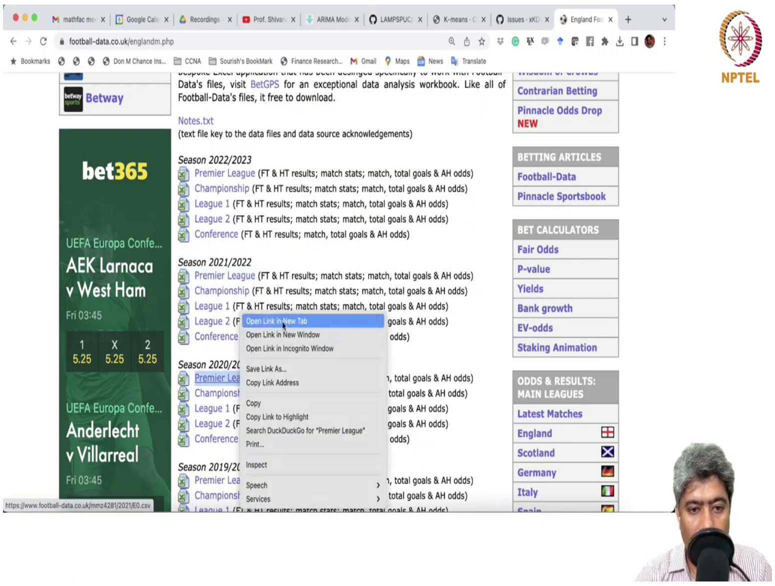
And, yeah.

(Refer Slide Time: 03:28)



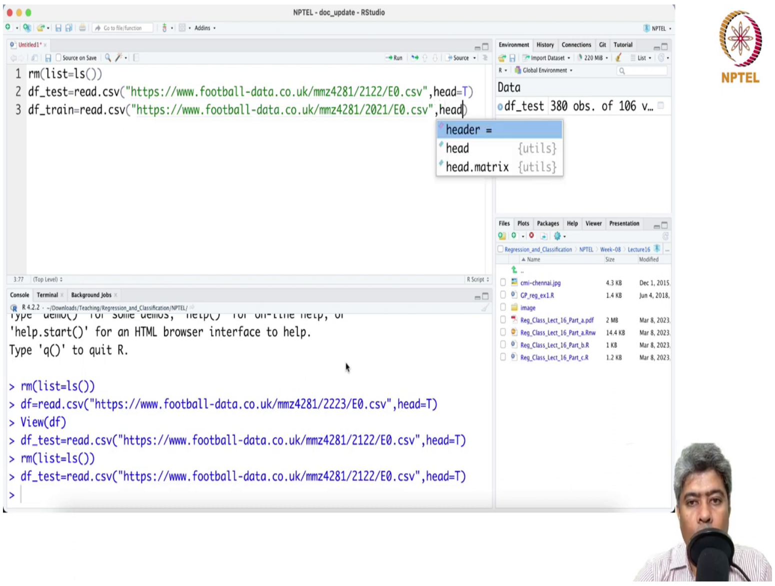
Now, and then df train sorry, train equal to read dot csv.

(Refer Slide Time: 03:42)



And then we take 20, 21 data, copy the link address.

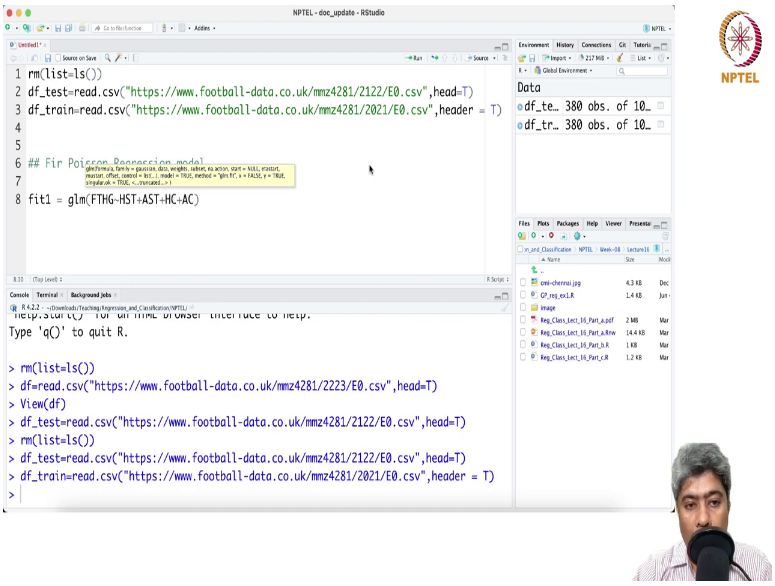
(Refer Slide Time: 03:49)



And head equal to true, ok. So, so the 2021 series data we will use as a train and 21, 22 series we will use as a test and we will see what is happening in that. So, the first what I will do. So, as we have explored this data a little bit before in a previous hands on. So, we know that we have built the Poisson regression model.

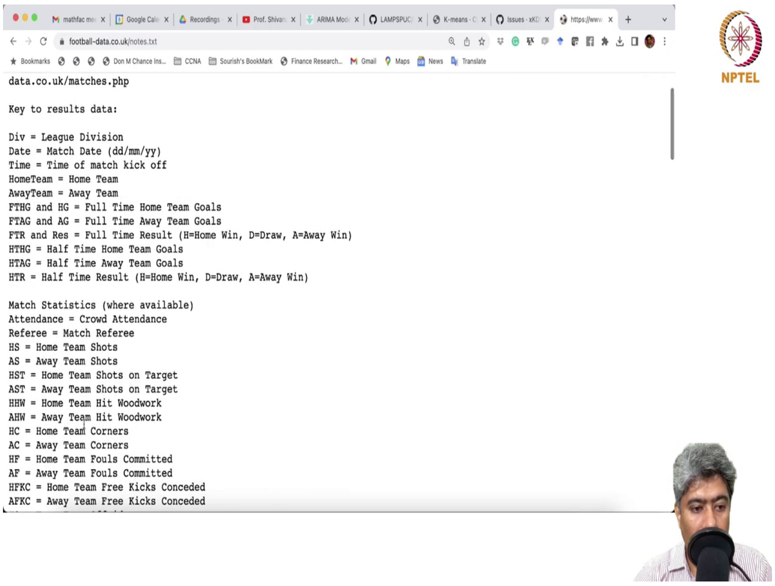
Let us first fit the Poisson regression model and then we will compare the performance of the Poisson regression model with the other models with the regression tree and the random forest.

(Refer Slide Time: 04:36)



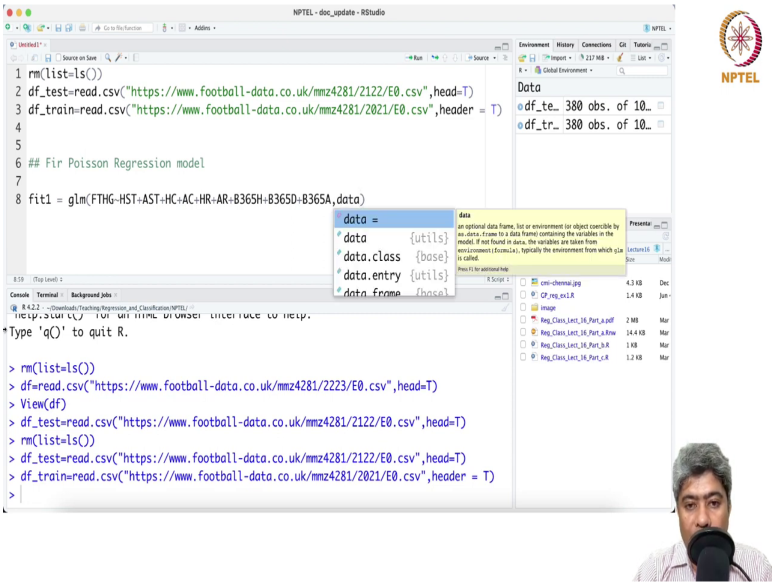
So, first fit Poisson Regression model, ok. So, fit1 equal to glm FTHG is the target variable full time how many goals scored by the home team and then HST how many shots are on target by the home team? AST how many shots are on target by the away team? HC, AC if we actually go and check the notes.

(Refer Slide Time: 05:30)



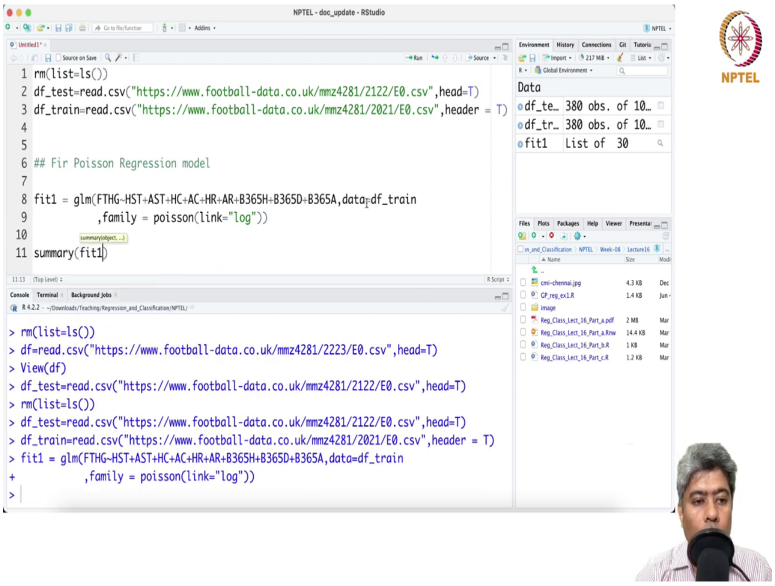
HC is home team corners, away team corners. And then actually we can put many more things. For example, shots target yeah.

(Refer Slide Time: 05:55)



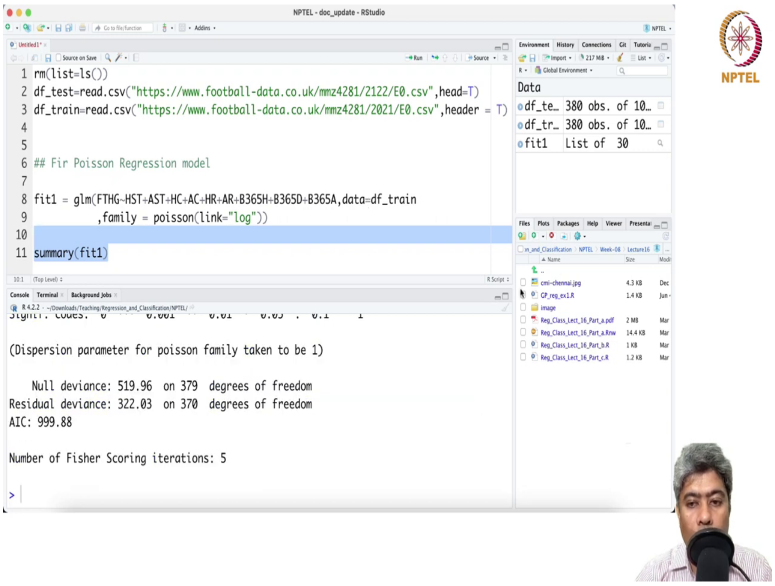
And then HR plus AR. HR was stands for home team red card, AR stands for away team's red card, HO and off side was not there, ok. And then bet we can also take the bet 365 if you go a little bit below. So, B365H plus B365 draw plus B365 away teams thing. So, this is a betting odds by bet 365 or house, betting house and we are going to use the train data set, ok.

(Refer Slide Time: 06:50)



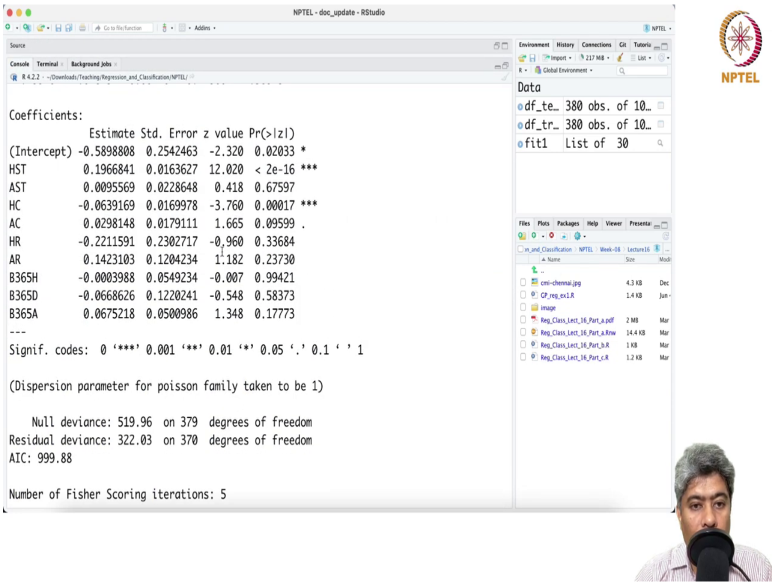
Data equal to df train and family, family equal to Poisson link equal to log, ok. So, if we run this and summery equal to fit1.

(Refer Slide Time: 07:13)



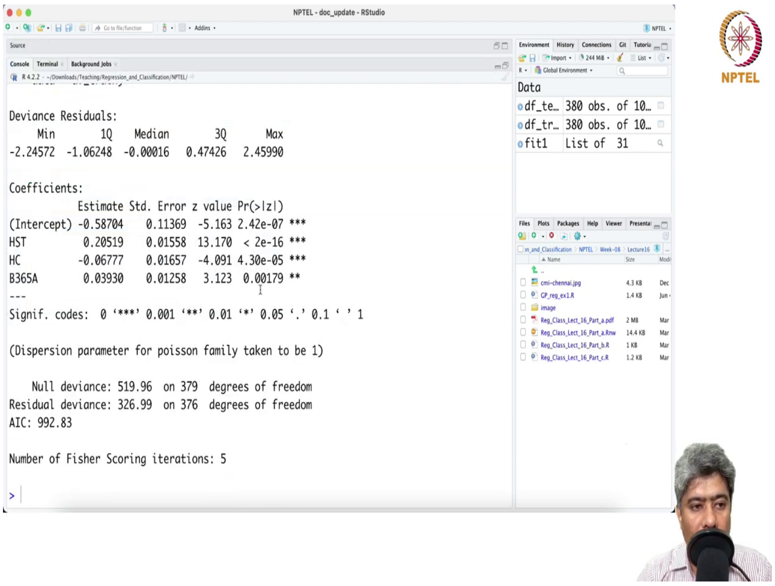
If you run this.

(Refer Slide Time: 07:15)



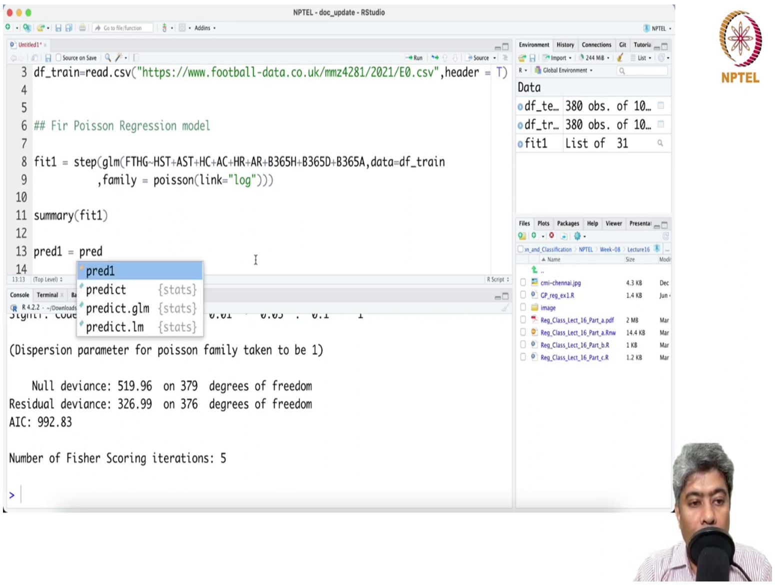
So, what we are seeing that home team shot made by the home team and home team how many corners they have made has a significant effect and rest we are not seeing many effect. One possibility could be one possibility could be there are multi collinearity issue. We can, but what we can do?

(Refer Slide Time: 07:44)



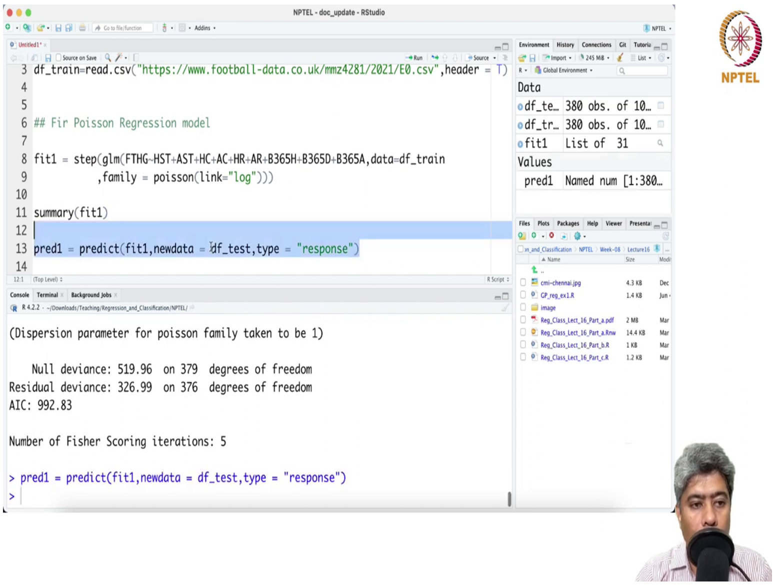
Before that we can just do a stepwise selection method to get the best model. So, if we just get a sort of a based on AIC. So, yeah, now, what is going here? That along with home team number of shots on target by home team and number corner by the home team we are also getting B365A the; that means, what is the away teams bet by B365 betting ratio odds that also playing a role, alright. So, we can we have some model. Let us take this is the model we have.

(Refer Slide Time: 08:32)

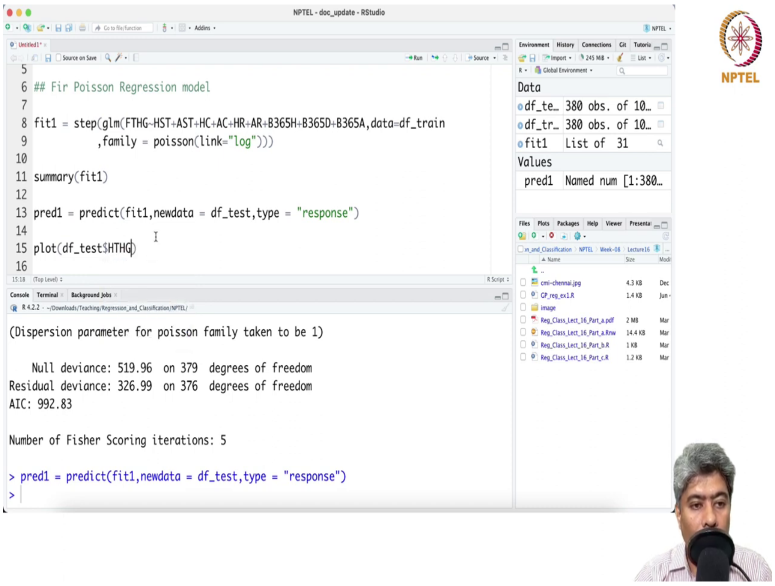


Now, what we can do? We can make the prediction for prediction for the test data set the next seasons data set pred predict predict df fit1 fit1 comma new data equal to df test, df test and type equal to type equal to response, ok.

(Refer Slide Time: 09:06)

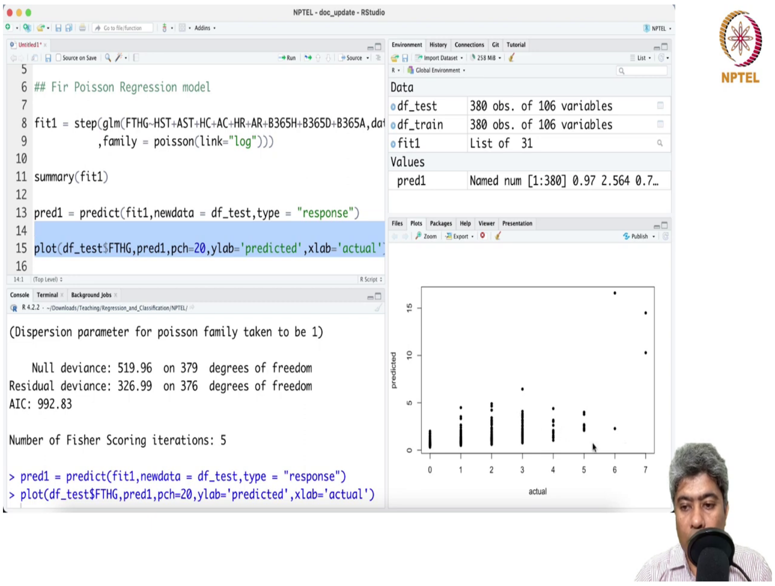


(Refer Slide Time: 09:24)



Now, what I can do I can plot the from the df test what is the HTHG or sorry, FTHG full time goals scored by home team and predicted values through predicted values.

(Refer Slide Time: 09:38)

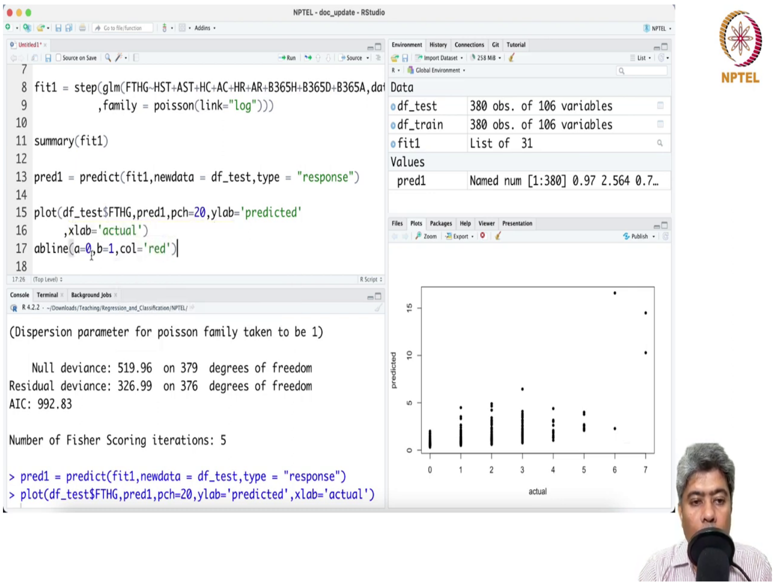


So, pch equal to 20, pred equal to ylab equal to predicted predicted and x lab x lab equal to actual.

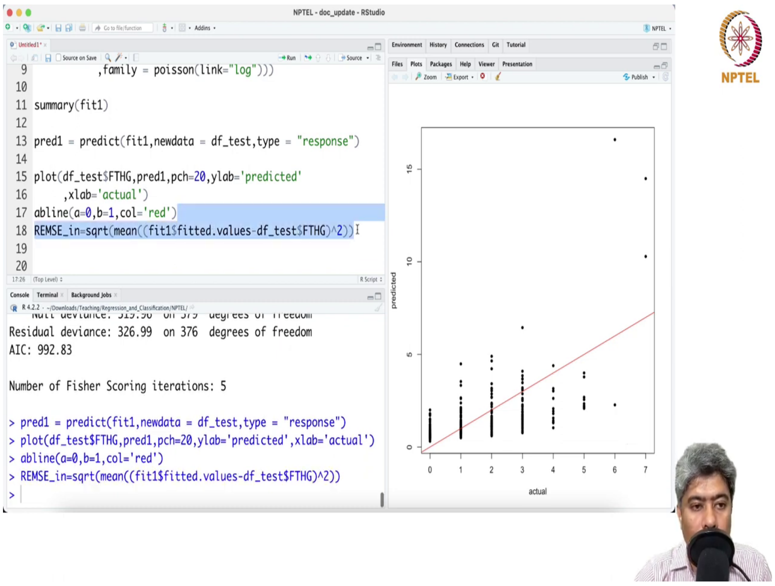
So, what we are seeing that actual was 7 you know somewhat it has some power, but and also there are cases where they predicted 15 goal, 10 goals it is little bit over estimating and actual was it was 7. So, the model has bit of a over estimating feeling like you know here there are cases where it has predicted 5 goals, but most of the goals were on the 3 goals.

So, it is bit of a predicting over estimating the bias, there is some looks like. So, one possible way is also we can do a abline a equal to 0 and b equal to 1 we put like this then color equal to red that also.

(Refer Slide Time: 11:03)



(Refer Slide Time: 11:18)

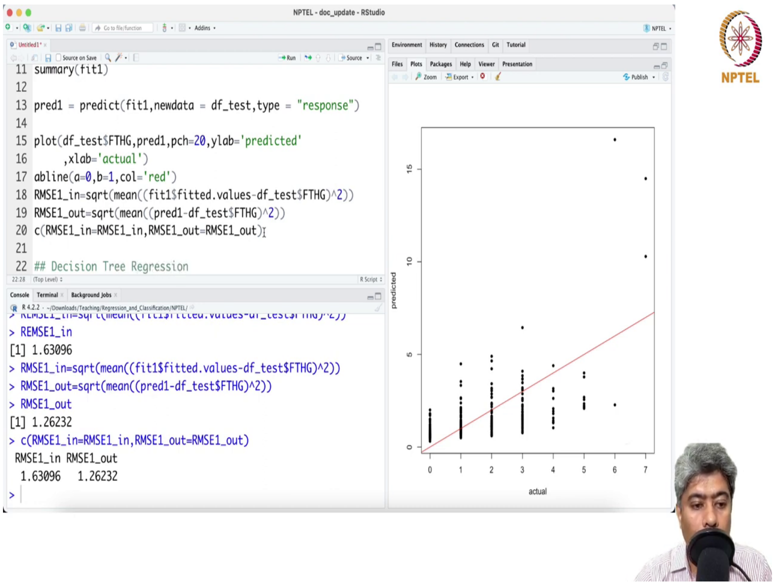


So, yeah. So, looks like some over estimation some under estimation here there are some under estimations, there are these places there are over estimations are happening, ok. So, this is what we are seeing. Now, what is what we are seeing here is the let us calculate the RMSC because one of the problem you know Poisson regression and the Poisson regression is a statistical model whereas, regression tree random forest are machine learning model for regression.

Now, what happens in the machine learning models there in these machine learning models there is no there are no likelihood you cannot write down the likelihood. Since you do not have a likelihood you cannot calculate AIC, BIC for these decision tree models, random forest models. So, you cannot use AIC, BIC type criteria to compare the random forest or decision tree against Poisson regression.

So, what we can do? We can use root mean square error which can be used for both Poisson regression as well as for the regression tree and random forest. So, we are going to calculate root mean square error for Poisson regression. So, RMSE in equal to. So, first from the fit1 we have to take the fitted values out minus from the df test dollar FTHG square it, take mean and then take the square root this is in sample root mean square error this is for RMSE1, ok.

(Refer Slide Time: 14:14)

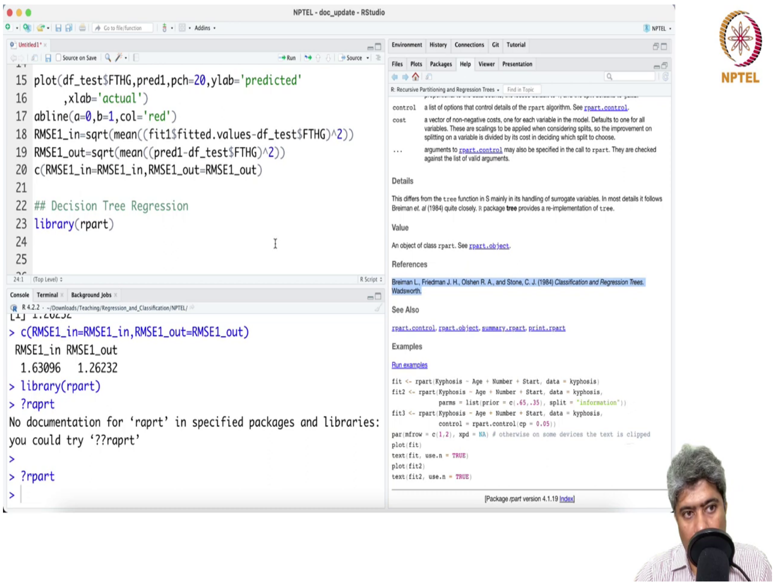


And then sorry, I think I have RMSE1 and then RMSE1 for out sample what I have to do I have to basically instead of this I have to take the predicted value I already predicted it I have to just go this and which is 1.26.

So, in sample there is lot of issues, but in out sample it looks like it is better that is sometime that is bit weird. Because typically in samples are better than the out of the sample, but we will see how it is doing in the yeah, we will see how it is doing in the regression t let us check how it how it is doing in the regression t and random forest decision tree and random forest.

So, next we will compare the decision tree regression. We will fit Decision Tree Regression.

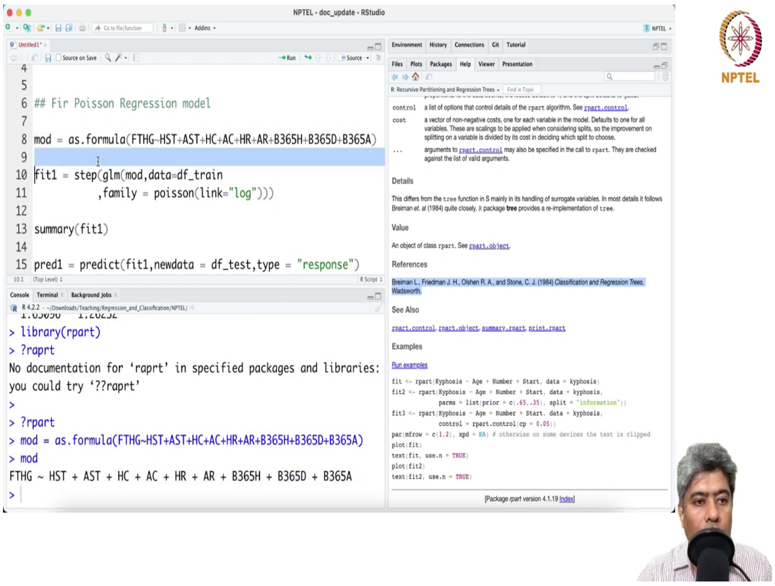
(Refer Slide Time: 15:28)



So, in the decision tree I am going to use the library called rpart ok, and if you come question mark rpart. So, it is called recursive partitioning and regression tree, ok. So, you just it is basically a regression tree it fits the essentially tree function in S mainly used to support it follows the Breiman's 1984 quite closely it says.

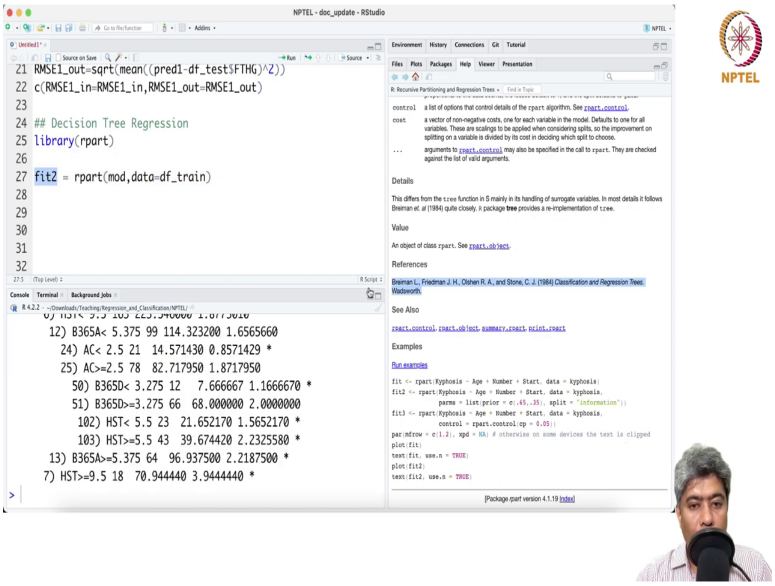
It most details most detail it follows Breiman et. al and in the main model main model proposed by Leo Breiman in 1984. So, rpart mostly follow that major thing this back paper classification and regression tree. So, this is being implemented in rpart. So, I am going to use the classical decision tree regression proposed by Leo Breiman implemented in R, ok.

(Refer Slide Time: 16:46)



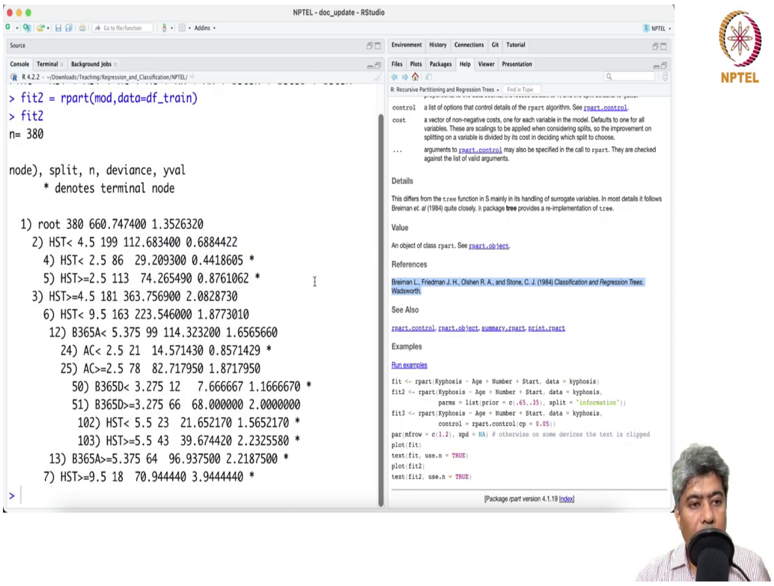
So, I am going to fit my second model the decision tree rpart. And then I am what I am going to do actually you know what? We can we are going to have the same model mod equal to at the rate formula sorry, s dot formula and I can have it and if I just run it here I think that will be, yeah.

(Refer Slide Time: 17:33)



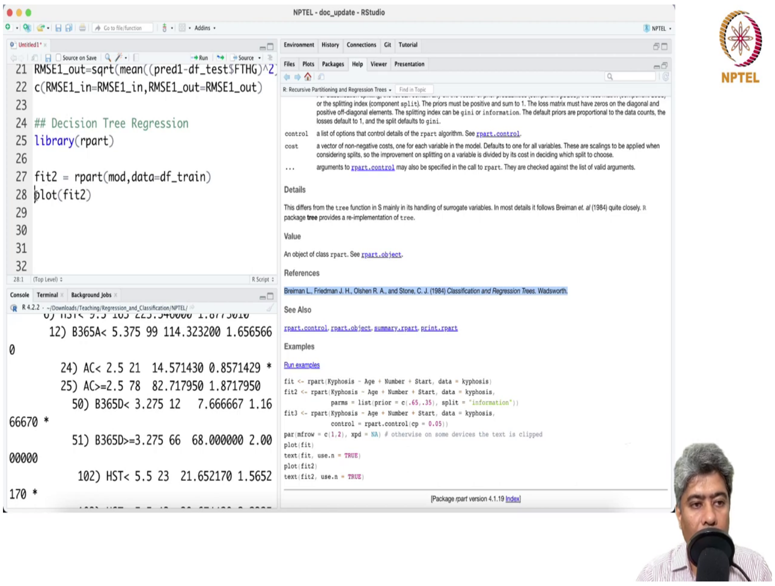
I think it is working fine. So, I will just call this model here, ok. If you run this model you can see this is the model I am going to fit and data equal to train ok, df train data equal to df train, ok. Now, if you run this. So, if you now run it.

(Refer Slide Time: 18:07)



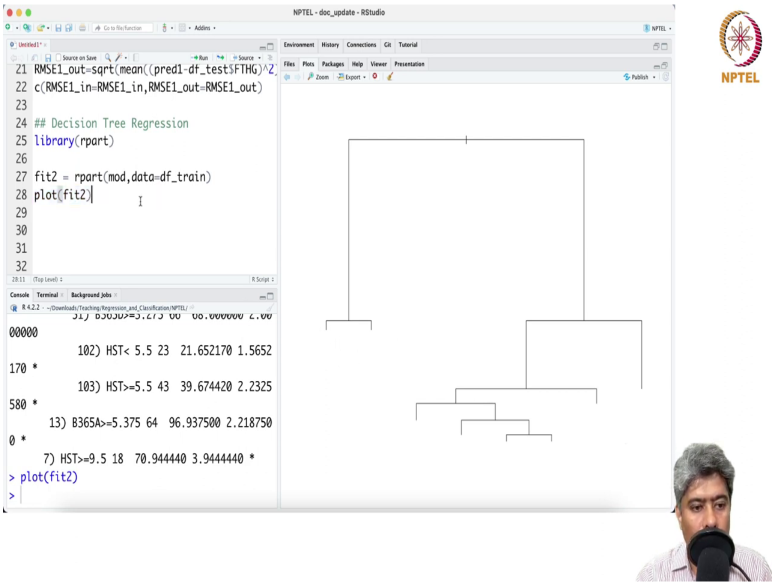
You will see that at the root it is saying what should be what it should does and then if HST equal to is less than 4.5 it should do something HST is less than 2.5, it should do something in this way. The decision tree is being made you can if you do plot fit2. So, generally it plot the decision tree nicely.

(Refer Slide Time: 18:41)



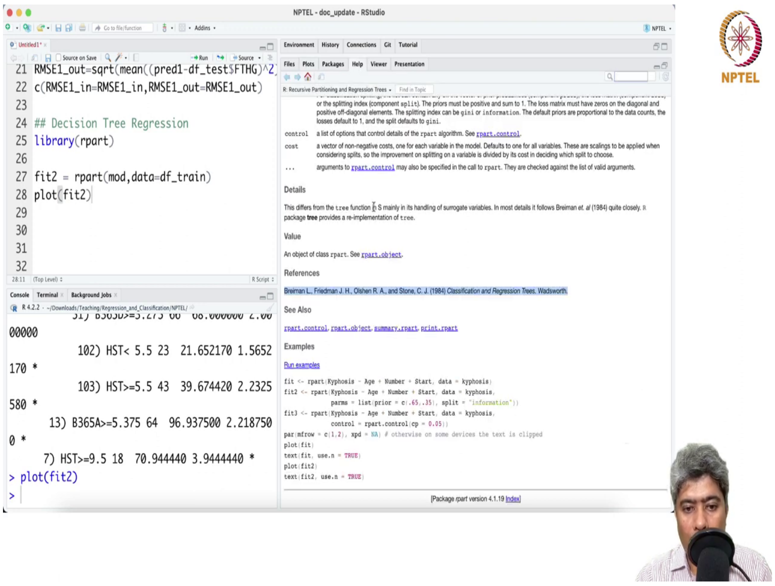
Let me try, let me try.

(Refer Slide Time: 18:43)



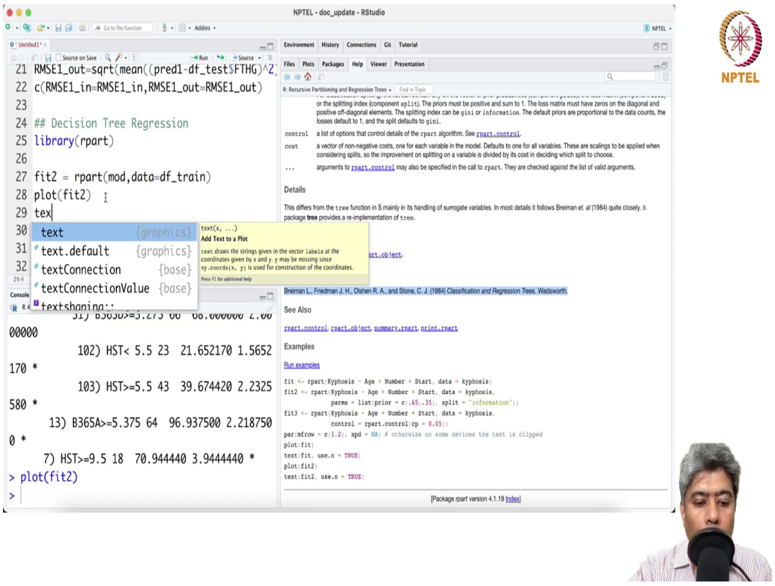
So, yeah, it is it does a decision tree, but for some reason it is not showing the plots.

(Refer Slide Time: 18:54)



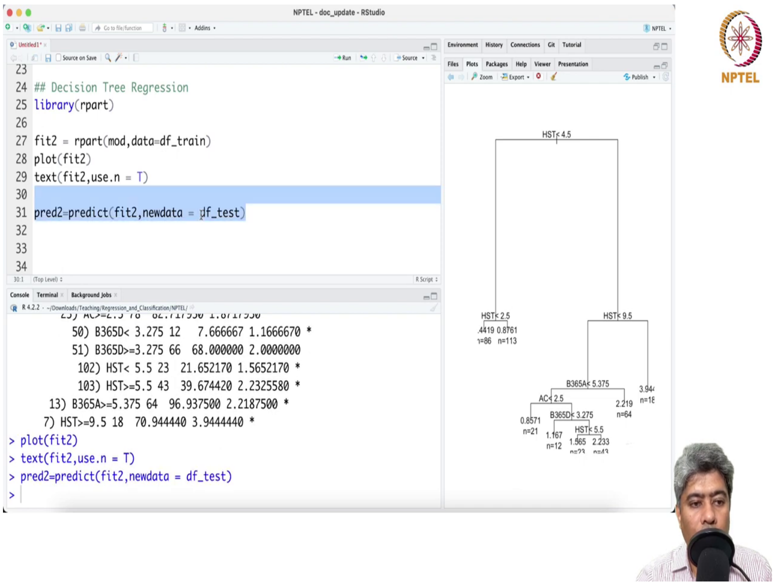
So, ok, it is and I have to say text use n dot true, alright.

(Refer Slide Time: 19:08)



So, I will say text.

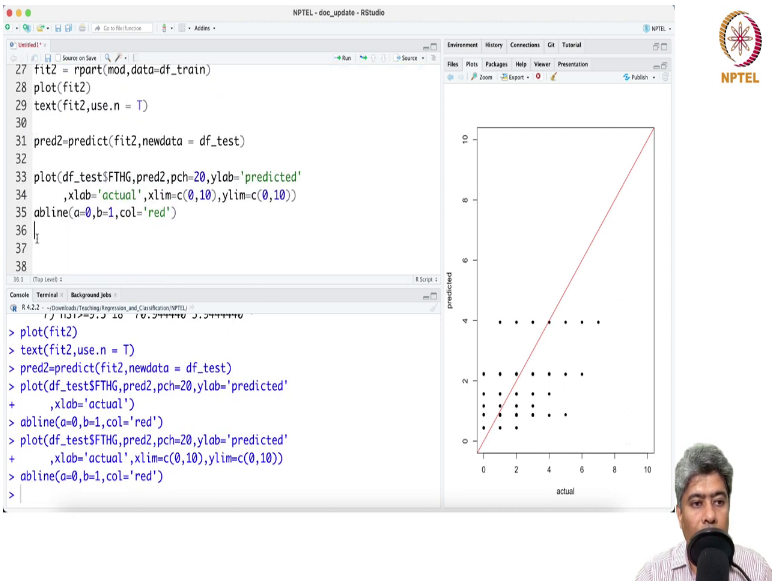
(Refer Slide Time: 19:11)



Text fit2 use dot n equal to true let me try, yeah, alright. So, it is saying that if HST is less than 4.5 it should go in this way otherwise it should go in this way. If it is less than 4 to 2.5 then with 0.4419 it take a decision it is I am not sure how this guys are doing these things, ok.

Let us do the prediction, let us do the prediction first then we will understand how it is doing. So, pred2 and pred2 equal to predict predict fit2 new data equal to df test, ok. If I just do that and now what I am going to do, I am going to just simply plot this guy with rate 2 and let us see how it is doing, ok.

(Refer Slide Time: 20:34)

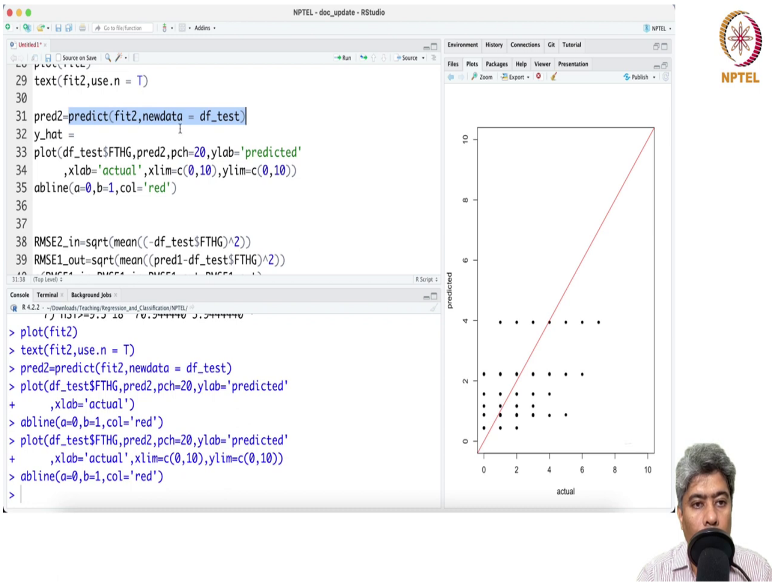


Interesting. So, the predicted values are somewhat constant you know for different 2 it is it is not much varying, but it is actually constant. And that is what we were saying like it should be constant, it takes within a region it takes a some bunch of constant values. But one interesting thing is it looks like it is it is not over estimating maybe it is little bit under estimating.

You can see that it is going somewhere between 0 the averages are going between half to 4 it is not making much differences. So, abline if we draw abline a equal to 0 and b equal to 1 with color equal to red, ok. So, yeah, it is not it is bit of a my feeling is it is bit of under estimating. Because you can see that x is taking range xlim is taking range between 0 to say 10 and ylim is taking if I just say the same scaling effect if we let us see if it what happens if it is gets how see, yeah.

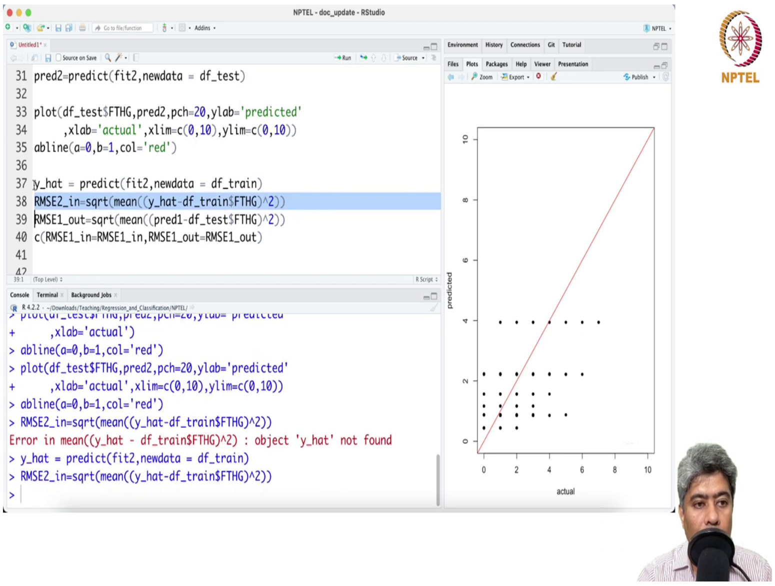
So, it is not even going up to the 6, 7, 8 whereas, in actual values are going up to this and it is it has a tendency to a little bit, it is trying to it is as a under estimation. So, that happens that is fine. Let us do some calculate the RMSE for in sample and out of the sample, ok. So, RMSE for the second model and instead of that what we have to do, ok.

(Refer Slide Time: 22:45)



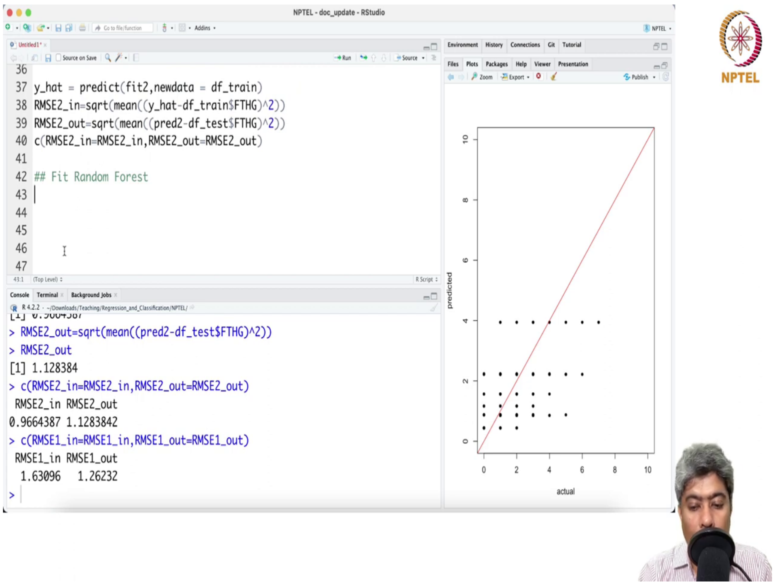
So, what I will do is red I will do a y hat here, y hat and fit to instead of test I will do train, ok. Let me just here and then y hat will be here and instead of test I have to take train FTHG, I have to first run the y hat of course.

(Refer Slide Time: 23:13)



And then this and then RMSE2 out pred2 and this now this out is 1.12. So, RMSE2 in and RMSE2 out is 2 out, ok. So, 0.96 and 1.12 whereas, we are.

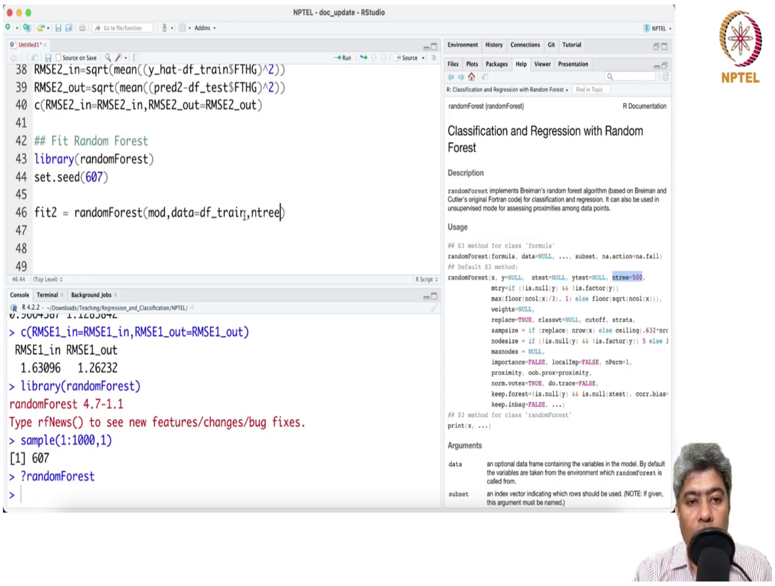
(Refer Slide Time: 23:52)



It was quite improvement if you see in sample Poisson was very high, but in sample is doing fit in out of the sample Poisson is doing still slightly better 1.26 whereas, oh no, decision tree is doing better because it is going to 1.128 which is smaller than 1.262. So, decision tree is doing better than the Poisson regression, ok.

So, now finally, finally, we will do, but we can see a little bit over estimation as well because in sample RMSE is 0.966 whereas, out of the sample is 1.12, it is slightly over estimation is happening. But in Poisson regression in sample is very high and out of the sample is bit low. So, that is also bit of a weird I do not see why it is happening, but let us see what is there, ok. So, let me fit the random forest, random forest.

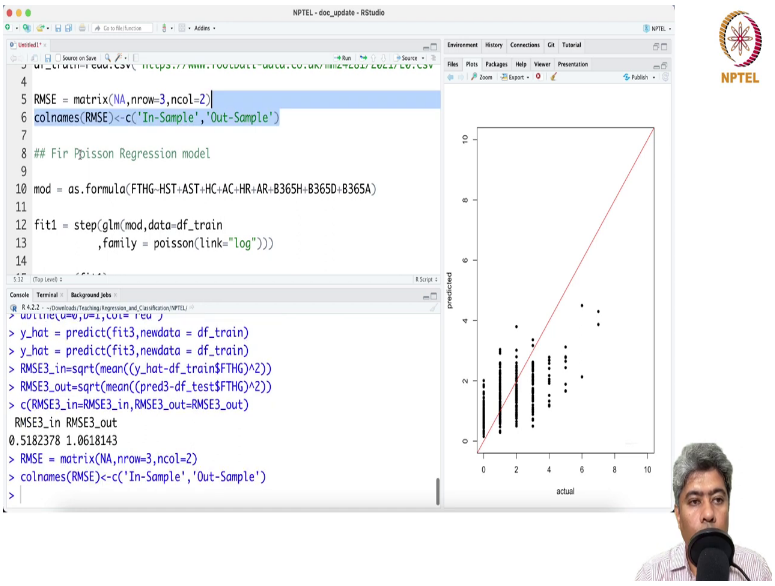
(Refer Slide Time: 25:17)



So, what I will do first I will call the library that is of library called randomForest, ok. And now since it will do sort of random sampling of the things that we will do sample 1 is to 1000 comma 1. So, 607 this is the set I am going to use.

And then what I am going to do, I am going to fit the third model instead copy this instead of that randomForest mod hm. Of course, one more thing I have to check few things randomForest. So, it is saying y NULL x test y test and then entry is number of trees that you have to give. So, maybe I will give 1000 and see fit3.

(Refer Slide Time: 26:32)

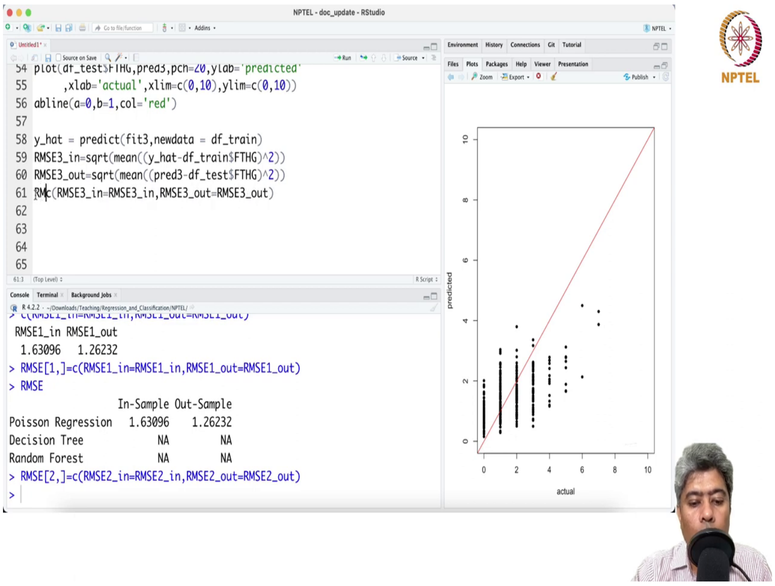


This is my third model and there are some other things are also there. So, what I will do, we will let us just keep it in there and ok, it does very good job if you fast it calculate things very fast. So, what I am going to do? I am going to copy this and first thing I will do is predict the my predict 3, ok. And let us see how predict does, ok. So, predict 3, ok.

So, looks like also some kind of under estimation because in the eighth prediction is in the out of the sample prediction it is not so high, ok. So, at the higher level it is some kind of under estimation is happening on the model site. Then what I am going to do I am going to calculate the RMSE for the third model y hat equal to fit3; obviously, this and then whatever the th this is my third RMSE, this is the third RMSE with pred 3rd prediction this is 3 this is 3, 3 and 3, ok, nice.

So, clearly the random forest is winning hands down, ok. So, let me write it down. So, let me a do one thing, let me just define a matrix called RMSE, ok. RMSE equal to matrix, ok. And NA nrow equal to 3 ncol equal to 2 col names equal to RMSE equal to In sample In sample and Out sample, ok.

(Refer Slide Time: 29:28)



So, this is In sample this is Out sample and row names equal to RMSE. If is first model was Poisson regression comma second model is decision tree and the third model is what is that third model is random forest, ok. And now what I am going to do I am going to put that in the 1, ok. So, yeah, and then going to put that 2 comma equal to this and finally, RMSE3 comma equal to.

(Refer Slide Time: 30:59)

Now, if I just run you can see in sample it is dropping constantly out of the sample also it dropped and Random Forest has 1 hands down, though we are seeing quite a bit of over fitting in both Decision Tree and Random Forest because Out sample RMSE is higher than the In sample. But in out of the sample the root mean square error is significantly lower, ok.

When we plotting this, we are seeing a bit of a underestimation is going on in a conservative site, but random forest is site sort of winning hands down, alright. So, that is how typically you do compare across the statistical models and the machine learning models.

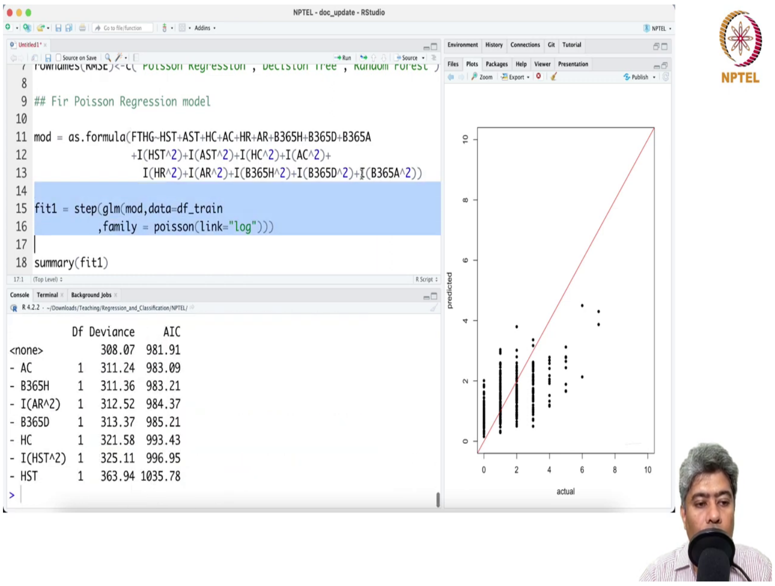
Remember that comparing Poisson Regression with Decision Tree and Random Forest is a bit of a unfair, because if you look into the model Poisson Regression model what I am doing is I am just fitting a linear hyperplane and I have not add any engineering in feature engineering.

If I add feature engineering perhaps possible that Poisson Regression model will start improving ok, whereas, Decision Tree and Random Forest the way the algorithms are being developed it will automatically try to fit the model in a way it will the algorithm will capture the non-monotonic non-linear behavior between the x and y in the higher dimension. So, this is a very strong very flexible model Decision Tree and Random Forest, Poisson Regression in that sense is a much more rigid conservative model.

So, in that sense it is not perhaps a very fair fight if you want to really apple to apple apple to apple thing comparison probably, we have to add few more lines like you know in feature engineering with HST square plus I and AST square plus I HC square plus I AC square plus I HR square plus I AR square plus I B365H square plus B365D square.

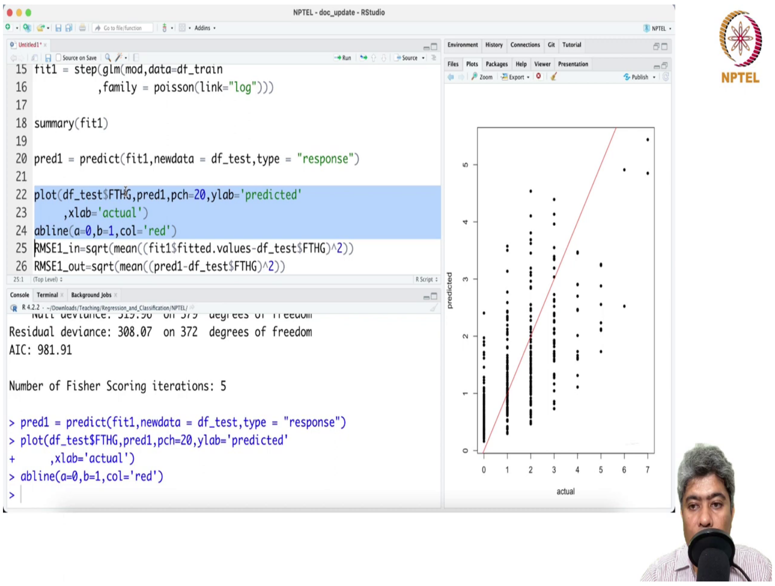
So, I am doing some feature engineering here just to check I do not know whether it will help the Poisson regression model or not, but you can try always try some feature engineering.

(Refer Slide Time: 34:26)



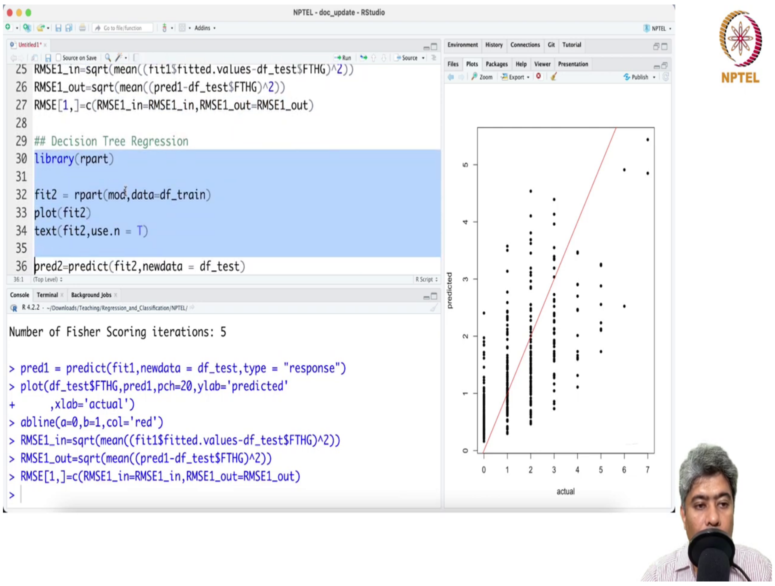
And if you run this and let us see summary. And you can see that some of the you know more features are now higher order features are now actually quite significant. Now, if you run this and let see how.

(Refer Slide Time: 34:50)



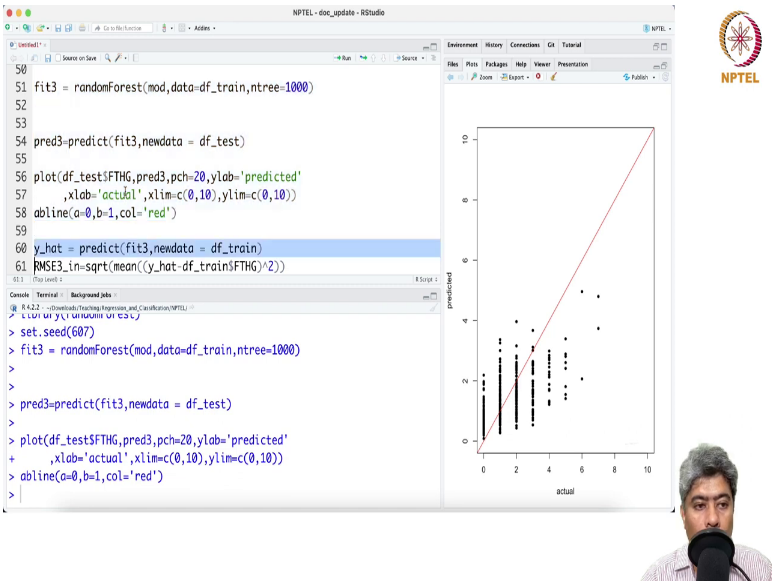
So, now it is looks like doing better.

(Refer Slide Time: 34:58)



If I run the same model and run me let us run these models, ok.

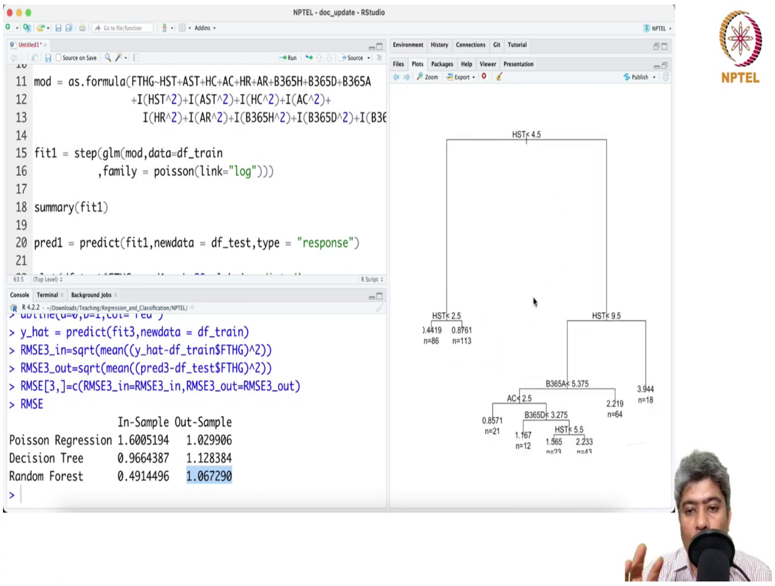
(Refer Slide Time: 35:18)



And now if I compare now, you see Poisson after adding the features Poisson Regression is doing even better than the higher order features, Poisson Regression is doing better than the Decision Tree. In fact, it is doing better than even Random Forest.

So, a simple Poisson Regression can do even better than machine learning model if you add the features engineered features correctly. So, sometimes simple and the advantage of Poisson Regression is you if the model is completely explainable, you can explain the model why it is happening. Well, Decision Tree also you can explain unfortunately this plot is not very coming up very nicely.

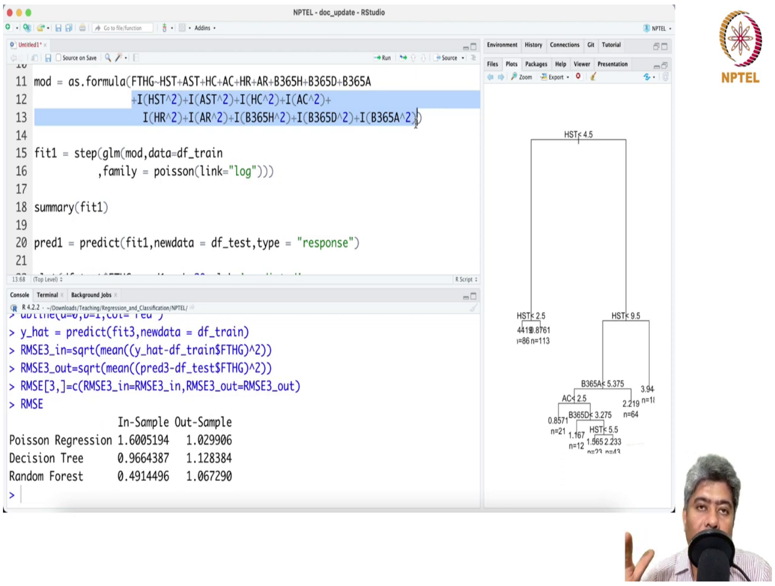
(Refer Slide Time: 36:15)



But overall, you can nicely explain what is happening here and, yeah. So, I think if you just let me zoom that, ok. So, what it is doing, it is basically saying that if HST is less than 4.5 and then you come by this route if HST is less than 2.5 then you will score a goal with rate 0.4419.

If it is greater than 2.5 then it you will score a goal with 0.8761, if it is greater than 4.5 then you come this side, if HST is less than 9.5 then you come this way, if the bit 365 is less than greater than 5365 5.37 then you will score a goal with rate 2.219, if it is greater than that then you will score a goal with rate 3.944. So, that is what it is saying effectively whereas. So, Decision Tree is also very good in terms of modelling, but you can see at the end.

(Refer Slide Time: 37:35)



When we add these engineered feature, when we add this engineered feature, ok. We found that Poisson Regression is doing even better than Decision Tree and random forest. But if you do not add this engineered feature then; obviously, it is just fitting a simple hyper linear line I mean linear hyper plane, basically straight line in two dimension and obviously, it is not doing very poorly compared to decision tree and random forest.

So, in my opinion you have to try all kinds of model with feature engineering and then you have to apply stepwise selection and dimension model dimension reduction technique and come up with a more parsimonious model and eventually you will see.

And effectively it is very difficult to say in my experience which model will be the best, there is no uniformly best model, in my experience you should try to fit all the models and see which model gives you the best fit. So, with this I will stop here and see you in the next week, next lecture.

Thank you very much see you.