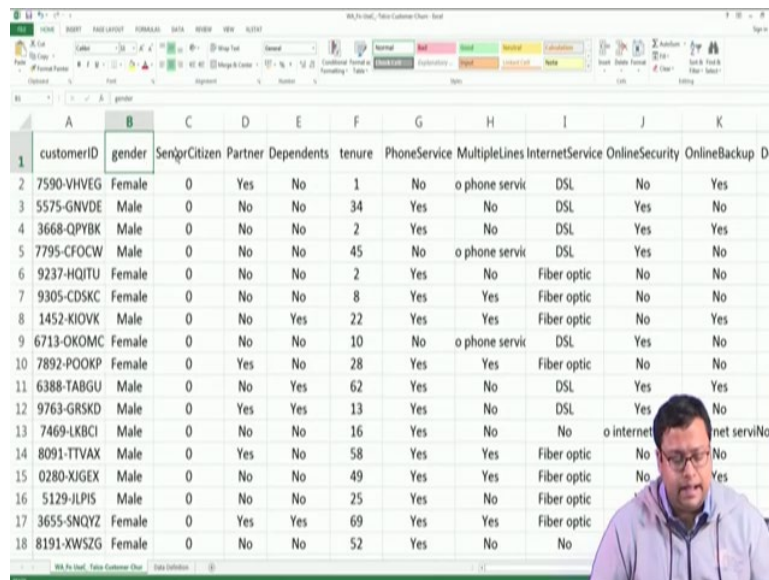


Marketing Analytics
Professor Swagato Chatterjee
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur
Lecture-46

Customer Churn and Customer Lifetime Value (Contd.)

Hello, everybody, welcome to Marketing Analytics course, this is Doctor Swagato Chatterjee from VGSOM, IIT Kharagpur who is taking this course for you. And in the last video we were discussing about customer churn prediction, I have given a very brief feedback about why customer churn is important. In this particular video, we will talk about how to predict churning behavior.

(Refer Slide Time: 0:38)



	A	B	C	D	E	F	G	H	I	J	K
1	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup
2	7590-VHVEG	Female	0	Yes	No	1	No	o phone servit	DSL	No	Yes
3	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No
4	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes
5	7795-CFOCW	Male	0	No	No	45	No	o phone servit	DSL	Yes	No
6	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No
7	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No
8	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes
9	6713-OKOMC	Female	0	No	No	10	No	o phone servit	DSL	Yes	No
10	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No
11	6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes
12	9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No
13	7469-LKBCI	Male	0	No	No	16	Yes	No	No	o internet	net servitNo
14	8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No
15	0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes
16	5129-JLPI5	Male	0	No	No	25	Yes	No	Fiber optic		
17	3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic		
18	8191-XWSZG	Female	0	No	No	52	Yes	No	No		

So why do not we open this week 9 S2.r file. This file is where the data is there and the corresponding data that we will be using is a Telco customer churn data. This has been downloaded from IBM's website, IBM Watson website, it is a freely available data set, is required to do academic work only. So, please use this data for that purpose. And we are opening this particular data, the data set looks like this. There, it is a big data, Let me explain first there are I think how many, around 7000 observations, so 7000 customers are there.

And if I further make it bigger you can see in the left side, the customer ID, then the gender and then whether this customer is a senior citizen or not, whether he has a partner or not, how many dependents, whether he has dependents or not, some demographic details about the customer is given. It has also been given that how much is his tenure, means how old is this customer in terms of months that whether the customer is very old or not.

(Refer Slide Time: 1:57)

	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	Streaming
1											
2	Yes	No	1	No	o phone serv	DSL	No	Yes	No	No	No
3	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No
4	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No
5	No	No	45	No	o phone serv	DSL	Yes	No	Yes	Yes	No
6	No	No	2	Yes	No	Fiber optic	No	No	No	No	No
7	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes
8	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes
9	No	No	10	No	o phone serv	DSL	Yes	No	No	No	No
10	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes
11	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No
12	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No
13	No	No	16	Yes	No	No	o internet serv	o internet serv	No internet serv	o internet serv	o internet serv
14	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes
15	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	Yes	Yes
16	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes
17	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes
18	No	No	52	Yes	No	No	o internet serv	o internet serv	No internet serv	o internet serv	o internet serv

	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBill
1										
2	o phone serv	DSL	No	Yes	No	No	No	No	nth-to-mo	Yes
3	No	DSL	Yes	No	Yes	No	No	No	One year	No
4	No	DSL	Yes	Yes	No	No	No	No	nth-to-mo	Yes
5	o phone serv	DSL	Yes	No	Yes	Yes	No	No	One year	No
6	No	Fiber optic	No	No	No	No	No	No	nth-to-mo	Yes
7	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	nth-to-mo	Yes
8	Yes	Fiber optic	No	Yes	No	No	Yes	No	nth-to-mo	Yes
9	o phone serv	DSL	Yes	No	No	No	No	No	nth-to-mo	No
10	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	nth-to-mo	Yes
11	No	DSL	Yes	Yes	No	No	No	No	One year	No
12	No	DSL	Yes	Yes	No	No	No	No	One year	No
13	No	No	o internet serv	o internet serv	No internet serv	o internet serv	o internet serv	o internet serv	o internet serv	No
14	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Yes	No
15	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Yes	Yes
16	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
17	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
18	No	No	o internet serv	o internet serv	No internet serv	o internet serv	o internet serv	o internet serv	o internet serv	No

Then if I come to this side, I have also whether the customer has phone service or not, whether there are multiple lines, how many phone services there, whether there are multiple lines or not. The internet service, what kind of internet service this customer has, whether it is a DSL or it has a fiber optic or it has no connection, those kind of things are there. Online security, if there is no internet service then there is no requirement of online security.

So there is yes, no and no internet service. Similarly online backup if it is no internet connection, there is no online backup is needed. No device production is needed, no tech support is needed, so these are yes-no questions. And streaming TV, streaming movies, whether they also consume those kind of things or not. So Telco is basically a telecom

company and they are trying to, telecom is a service company and they make money every month slowly.

So they are saying that whether these, which kind of services these particular customers consume. And for how many months they are consuming the services with me, this tenure is actually a how many months they are consuming the service for me. And then the contract, the contract is what kind of contract, is it a month to month contract, monthly basis or yearly contract. So, what mode of contract is another variable important. Paperless billing is there or not is another variable.

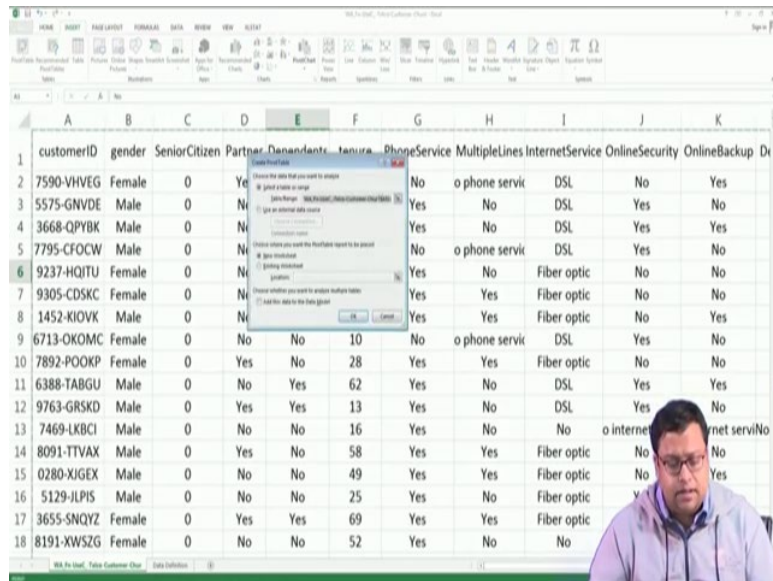
(Refer Slide Time:3:20)

	L	M	N	O	P	Q	R	S	T	U
	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
1										
2	No	No	No	No	nth-to-mo	Yes	Electronic check	29.85	29.85	No
3	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
4	No	No	No	No	nth-to-mo	Yes	Mailed check	53.85	108.15	Yes
5	Yes	Yes	No	No	One year	No	credit transfer (autom	42.3	1840.75	No
6	No	No	No	No	nth-to-mo	Yes	Electronic check	70.7	151.65	Yes
7	Yes	No	Yes	Yes	nth-to-mo	Yes	Electronic check	99.65	820.5	Yes
8	No	No	Yes	No	nth-to-mo	Yes	credit card (automa	89.1	1949.4	No
9	No	No	No	No	nth-to-mo	No	Mailed check	29.75	301.9	No
10	Yes	Yes	Yes	Yes	nth-to-mo	Yes	Electronic check	104.8	3046.05	Yes
11	No	No	No	No	One year	No	credit transfer (autom	56.15	3487.95	No
12	No	No	No	No	nth-to-mo	Yes	Mailed check	49.95	587.45	No
13	No internet service	internet servi	internet servi	internet servi	Two year	No	credit card (automa	18.95	326.8	No
14	Yes	No	Yes	Yes	One year	No	credit card (automa	100.35	5681.1	No
15	Yes	No	Yes	Yes	nth-to-mo	Yes	credit transfer (autom	103.7	5036.3	Yes
16	Yes	Yes	Yes	Yes	nth-to-mo	Yes	Electronic check	105.5	2686.05	No
17	Yes	Yes	Yes	Yes	Two year	No	credit card (automa	113.25	7895.15	No
18	No internet service	internet servi	internet servi	internet servi	One year	No	Mailed check	20.65	1022.95	No

The payment method whether he is paying through electronic check or whether he is paying through mail check or bank transfer or credit card, these are various kind of payment methods that they are using. And then there is monthly charges, total charges and churn. So, ultimately the prediction variable is churn, whether somebody is churning or not. So, first before I jump into predicting whether somebody churns or not, it is a very good practice to understand this data.

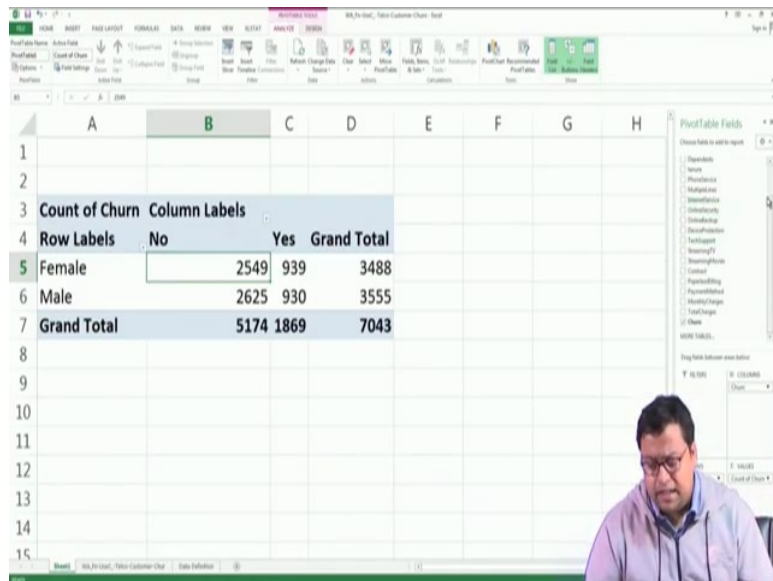
See, it is not important or it is not only important to predict who is going to churn but further more important for a marketing manager is how to stop churn. It is important to know that how I can do a simple thing which can stop churning. And that is why it is important to know that what are the predictors of churn, which are the variables which will lead to churning or not. So, for example, if I can see that whether there is a probability...

(Refer Slide Time: 4:25)



So, what I am doing is I am doing a pivot analysis. So I am selecting any cell here, going to insert pivot table, I am clicking on this pivot table, and it selects the range and pressing Okay, so now I have the pivot table.

(Refer Slide Time: 4:48)



And here in the right side, I am dragging let us say churn as my values or churn as my columns and let us say gender as my rows and then count of churn as my values also. So if you can see carefully that out of males, 3555 males in this data set, I will just make it bigger out of 3555 males 930 are yes, 2625 are no. And here 939 are yes and 2549 is no. So I do not think there is much difference in terms of male and female and churning and not churning. So I do not think gender is one of the major predictor of churning.

(Refer Slide Time: 5:43)

Count of Churn	Column Labels		
Row Labels	No	Yes	Grand Total
0	4508	1393	5901
1	666	476	1142
Grand Total	5174	69	7043

So this kind of basic initial level analysis you can do. Or that is senior citizen, so if I can check. Okay, so senior citizen, you will be amused to see that if you are a senior citizen, your chances of churning is very high, almost 50%. If you are a not senior citizen, then your chances of churning is very low, almost like less than 25%. So this is something which is interesting and counter intuitive.

So we had an idea that the senior people will be more sticky towards a service provider and younger persons who might have more connectivity, who might have more information available with them, they are more tech savvy, they will have a tendency of switching. But here we are not seeing that we are seeing the opposite thing, we are seeing that senior people, out of 1142 senior people, 1100 senior people 476 is willing to switch, so that is almost 40% is willing to switch.

On the other hand 5901 out of 6,000 people, around 1400 people are actually willing to switch which is less than 25%. So, this is counter intuitive. Okay. Similarly, let us say partner.

(Refer Slide Time: 6:58)

Count of Churn	Column Labels		
Row Labels	No	Yes	Grand Total
No	2441	1200	3641
Yes	2733	669	3402
Grand Total	5174	1869	7043

So partner, I think that people who has a partner, who does not have partner are less switch prone, who has partner or more switch prone. There has to be some reason of that. Probably, so you have to actually discuss with yourself that why, as a marketing manager you have to discuss that why having a partner or not having a partner will actually impact your decision. Because these might be more sticky because your partner and you might be using the same telecom company, there can be some offers for your within network calls, within network message services and etc.

Sometimes you also share the same data service with the company. So the chances of switching if you have a partner is generally comes down. On the other one switching of, switching a telecom company when there is no partner probably will be that is why much higher.

(Refer Slide Time: 7:50)

Count of Churn	Column Labels		
Row Labels	No	Yes	Grand Total
No	3390	1543	4933
Yes	1784	326	2110
Grand Total	5174	1869	7043

What about dependents? See, it is the same thing the partner and dependent stand goes hand in hand. So, if you have dependents, there are lots of guys who are, so if you have lots of kids then you might be using the same one common internet service, one common data service, backup services and etc, sometimes streaming services also, TV streaming and so on. So then the chances of switching comes down but if you are a single person, then your chances of switching will go up.

(Refer Slide Time: 8:36)

Count of Churn	Column Labels		
Row Labels	No	Yes	Grand Total
0-19	1574	1233	2807
20-39	1095	320	1415
40-59	1121	217	1338
60-79	1384	99	1483
Grand Total	5174	1869	7043

Now, if you talk about tenure, so tenure has been taken here. So tenure I cannot see it like this because it is a continuous data, but I can break it the tenure between let us say, if I just put it here, and group tenure, let us say I am grouping tenure, right click group tenure by let us say,

0 to 72, I am grouping it by 20. So you will see when in the tenure is 0 to 19 the chances is high switching and slowly as the tenure increases the chances of this thing comes down, chances of switching comes down.

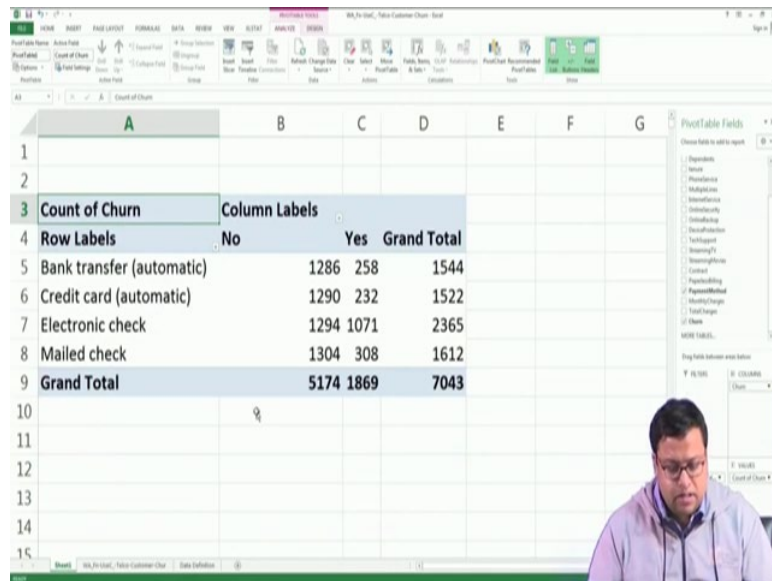
So 320 by 1000, this is one third and 200 by, it is one fifth and 99 by this thing is so much one thirteenth. So, this is how slowly the chances of switching, chances of churning will come down, when the tenure goes up.

(Refer Slide Time: 9:45)

Count of Churn	Column Labels		
Row Labels	No	Yes	Grand Total
18.8	1		1
18.85	1	1	2
18.9	1		1
19	1		1
19.05	1		1
19.1	2	1	3
19.15	1		1
19.2	4		4
19.25	2	1	3
19.3	2	2	4
19.4	2	1	3

Then if I just say that monthly total charges, how much you charge and if I then group this one so, it is saying that I cannot group this one, I am not sure why because it is continuous data I think. But even then if I just check this thing, we will see that as we go up the charges, we have to see that how the charges and this thing are related to each other, we have to run a correlation analysis here instead of this kind of grouping and we have to see that whether that works here as well or not.

(Refer Slide Time: 10:30)



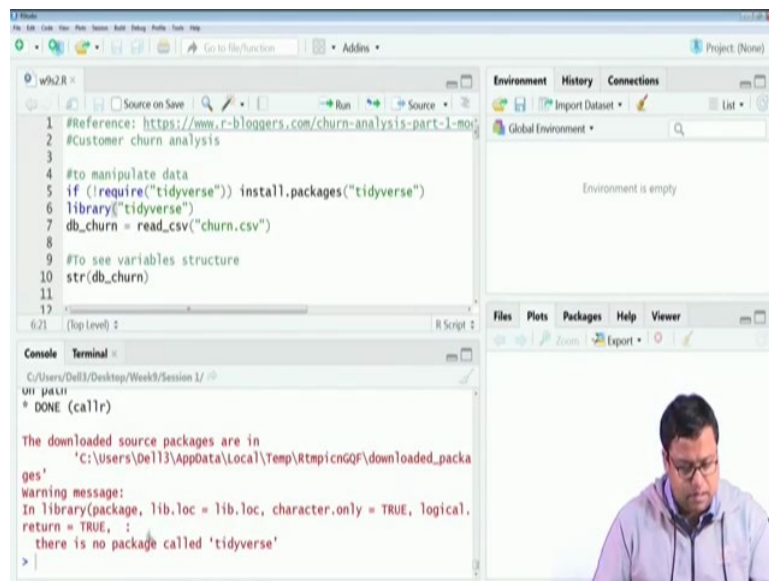
The screenshot shows an Excel PivotTable with the following data:

Count of Churn	Column Labels		
Row Labels	No	Yes	Grand Total
Bank transfer (automatic)	1286	258	1544
Credit card (automatic)	1290	232	1522
Electronic check	1294	1071	2365
Mailed check	1304	308	1612
Grand Total	5174	1869	7043

So, yeah so, we cannot do that here. So, this will not work but let us say PayPal is billing, a payment method. So, payment method wise also I can say that electronic cheque guys, those who are electronic, giving data in electronic cheque that means they are more I would say tech savvy than who mail cheque or who are doing bank transfer or credit card. So, these guys are more concerned towards switching and these guys are less switching prone. So this kind of data we are getting here.

Now what I have done, so description of the data is given here you can check, the senior citizen 1 is equal to yes, 0 is equal to no. So, that description is there. So, I do not save this data, what I did is I have changed this particular data to this churn data and I will try to predict chart.

(Refer Slide Time: 11:20)



The screenshot shows the RStudio interface. The main editor window contains an R script with the following code:

```
1 #Reference: https://www.r-bloggers.com/churn-analysis-part-1-mo
2 #Customer churn analysis
3
4 #to manipulate data
5 if (!require("tidyverse")) install.packages("tidyverse")
6 library("tidyverse")
7 db_churn = read_csv("churn.csv")
8
9 #To see variables structure
10 str(db_churn)
11
12
```

The console window shows the output of the script:

```
C:\Users\Del13\Desktop\Week9\Session 1/ >
> [1]
* DONE (callr)

The downloaded source packages are in
 'C:\Users\Del13\AppData\Local\Temp\Rtmp1cngQF\downloaded_packa
ges'
warning message:
In library(package, lib.loc = lib.loc, character.only = TRUE, logical.
return = TRUE, :
there is no package called 'tidyverse'
```

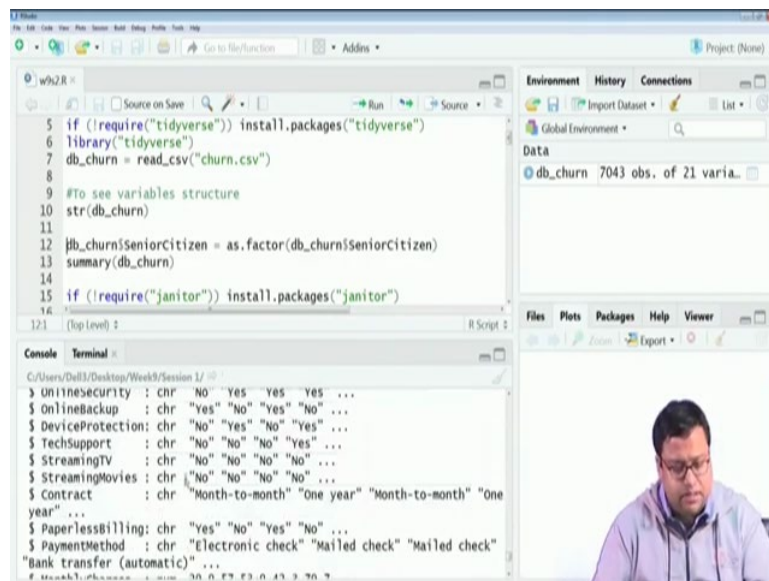
So, the code has been taken from Rbloggers.com it has been and the link has been shared here. So, we have changed the code depending on our necessary things whatever we are trying to do, but it has been, I would say inspired by this particular link. So, you might get some number of similarities there. So, what is, the first thing that I am to do is to check that if tidy verse is there, if tidy verse is there I will install the package tidy verse.

So I will just run this line, it checks whether tidy verse is there or not. If it is not there, it will install this particular library. So it might take one minute, so I will wait for 1 minutes or so, to let this particular thing get installed. So, tidy verse will be required to do certain analysis here. So, let us install it. So, while installing you see that the next thing, the next step is I am calling this library here and then I will be reading the data.

So, I have to set working directory to source file locations. So and then I have to read this data, some of the steps after this I will be doing quickly because I am just explaining it right now. So, this is quite quickly installing all these libraries. And once the libraries are installed, we will be calling this particular library, these are dependencies basically. Dependencies mean some of the functions of this particular libraries will be used while doing this, while using tidy verse.

So it is installing, still installing, it is almost done and now it will be done I think. Yes, so it is done very good. So, there is a warning message, well, there is no package called tidy verse, that is fine.

(Refer Slide Time: 13:16)

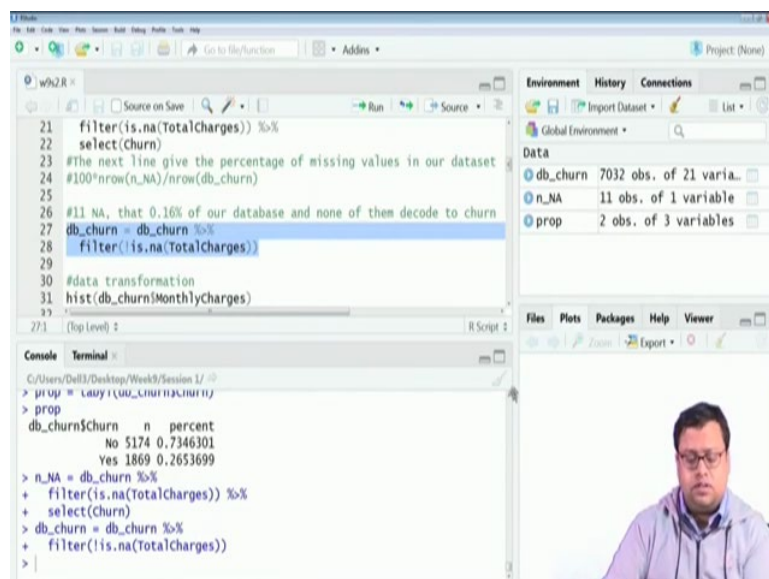


```
5 if (!require("tidyverse")) install.packages("tidyverse")
6 library("tidyverse")
7 db_churn = read_csv("churn.csv")
8
9 #To see variables structure
10 str(db_churn)
11
12 jdb_churn$SeniorCitizen = as.factor(db_churn$SeniorCitizen)
13 summary(db_churn)
14
15 if (require("janitor")) install.packages("janitor")
```

```
C:/Users/Dell/Desktop/Week9/Session 17 >
> unlinesecurity : chr "No" "Yes" "Yes" "Yes" ...
> OnlineBackup : chr "Yes" "No" "Yes" "No" ...
> DeviceProtection: chr "No" "Yes" "No" "Yes" ...
> TechSupport : chr "No" "No" "No" "Yes" ...
> StreamingTV : chr "No" "No" "No" "No" ...
> StreamingMovies : chr "No" "No" "No" "No" ...
> Contract : chr "Month-to-month" "One year" "Month-to-month" "One
year" ...
> PaperlessBilling: chr "Yes" "No" "Yes" "No" ...
> PaymentMethod : chr "Electronic check" "Mailed check" "Mailed check"
"Bank transfer (automatic)" ...
# Month:prop: : chr 30 0 57 13 0 15 3 70 ?
```

So, I will call this library now, tidyverse, and it will be, once it is called we will read the data. So yes, so, this is done and there are some issues, but there is no errors. And once that is done, I will save session data into source file location and read the data. So the data has been read. And then if I check the structure of the data, the structure of the data is sourced like this, that these are all character variables. So, if these are, these are all character variables that I am getting the phone service, the multiple lines, the internet service, online security, online backup, I have to convert them to factor.

(Refer Slide Time: 13:58)



```
21 filter(is.na(TotalCharges)) %>%
22 select(churn)
23 #The next line give the percentage of missing values in our dataset
24 #100*nrow(n_NA)/nrow(db_churn)
25
26 #11 NA, that 0.16% of our database and none of them decode to churn
27 db_churn = db_churn %>%
28 filter(!is.na(TotalCharges))
29
30 #data transformation
31 hist(db_churn$MonthlyCharges)
```

```
C:/Users/Dell/Desktop/Week9/Session 17 >
> prop = summary(db_churn$churn)
> prop
db_churn$churn n percent
No 5174 0.7346301
Yes 1869 0.2653699
> n_NA = db_churn %>%
+ filter(is.na(TotalCharges)) %>%
+ select(churn)
> db_churn = db_churn %>%
+ filter(!is.na(TotalCharges))
> |
```

So the first thing is I convert the senior citizen to factor and then I also use a janitor library and the library called janitor. So, this will not take much time, this is a small library that will

be used. Yes and that has been called and then what I will do is I will just check the properties. And it is saying that in my data, there is 5174 no churns and 1869 churns, which is 26% of people are churning, they are leaving this particular service provider and 73% people are not leaving the service provider.

Now, using this tidy verse library I will be using this, that whenever there is a total charges I will be, so if it is not, other than total charges, everything I will convert to factor variable in this db churn data set. And then what I will do is I will check that how many missing values are there, based on that I will reduce the number of variables. So, I am just checking the missing values is na total charges. It is saying that in total charges, how many missing values are there and I am filtering it based on that.

So, whenever there is no missing value, take the data, whenever there is missing value, do not take that data. So, that kind of filtering I am doing. So, we will check this thing 7043 was the previous case, it has come down to 7031. So 7042 to 7031 is 11 and so, 11 missing data has been dropped now. Then what? Then I will just draw the histogram of monthly charges.

(Refer Slide Time: 15:44)

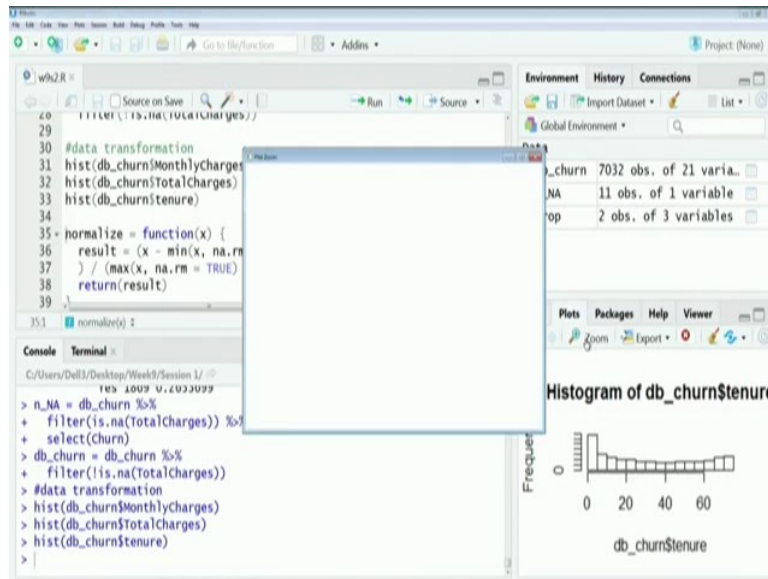
The screenshot displays the RStudio interface. The script editor on the left contains the following R code:

```
20  
29  
30 #data transformation  
31 hist(db_churn$MonthlyCharges)  
32 hist(db_churn$TotalCharges)  
33 hist(db_churn$tenure)  
34  
35 normalize = function(x) {  
36   result = (x - min(x, na.rm = TRUE)) /  
37   (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))  
38   return(result)  
39 }
```

The console at the bottom shows the execution of the code, including the output of the histogram function and the results of filtering missing values:

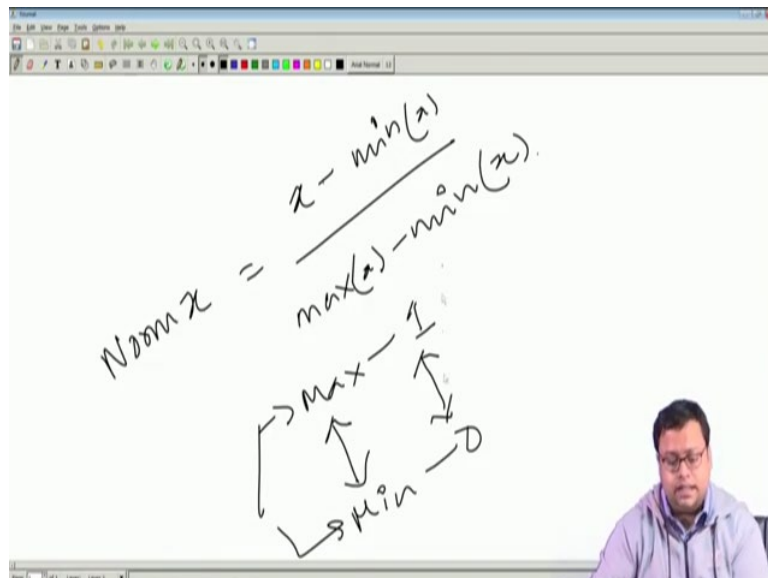
```
> n_NA = db_churn %>%  
+ filter(is.na(TotalCharges)) %>%  
+ select(churn)  
db_churn = db_churn %>%  
+ filter(!is.na(TotalCharges))  
#data transformation  
> hist(db_churn$MonthlyCharges)  
>
```

The Environment pane on the right shows the data frame 'db_churn' with 7032 observations and 21 variables. The Plots pane displays a histogram titled 'Histogram of db_churn\$MonthlyCharges' with a frequency axis from 0 to 100 and a monthly charges axis from 0 to 120. A small inset video of a person is visible in the bottom right corner of the RStudio window.



You see the histogram of monthly charges looks like this. So, which is not fairly distributed, which is not and this one is further skewed and tenure is also not equally distributed, it is also skewed. So, if that is the case, then I have to and the distributions are different in shape also. So, not everybody has the same shape. So, if that is the case, I have to go and normalize it, so the normalization function is, the output is basically $x - \min(x) / (\max(x) - \min(x))$

(Refer Slide Time: 16:33)



So, we have done this normalization before, Norm $x = x - \min(x) / (\max(x) - \min(x))$

So, that is something that is conversion I am doing, such that the maximum values takes the value of 1 and the minimum value takes the values of 0 and in between values take, in between values between 1 and 0. So, something like that we are trying to do here as well.

(Refer Slide Time: 17:07)

The screenshot displays the R Studio environment. The script editor contains the following code:

```
37 } / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))
38 return(result)
39 }
40
41 db_churn2=data.frame(db_churn)
42 db_churn2$TotalCharges=normalize(db_churn2$TotalCharges)
43 db_churn2$MonthlyCharges=normalize(db_churn2$MonthlyCharges)
44 db_churn2$tenure=normalize(db_churn2$tenure)
45
46 for(i in c(1:5,7:16))db_churn2[,i]=ifelse(db_churn2[,i]=="No internet
47 for(i in c(1:5,7:16))db_churn2[,i]=ifelse(db_churn2[,i]=="No phone st
48
49
50
```

The Environment pane on the right shows the following objects:

- Global Environment
- Data
 - db_churn 7032 obs. of 21 variables
 - db_churn2 7032 obs. of 21 variables
 - n_NA 11 obs. of 1 variable
 - prop 2 obs. of 3 variables
- Functions
 - normalize function (x)

The Console pane shows the execution of the following commands:

```
> normalize = function(x) {
+   result = (x - min(x, na.rm = TRUE)
+   ) / (max(x, na.rm = TRUE) - min(x, na.rm = TRUE))
+   return(result)
+ }
> db_churn2=data.frame(db_churn)
> db_churn2$TotalCharges=normalize(db_churn2$TotalCharges)
> db_churn2$MonthlyCharges=normalize(db_churn2$MonthlyCharges)
>
```

A histogram titled "Histogram of db_churn\$tenure" is visible in the bottom right corner, showing the frequency distribution of the tenure variable.

So we will be using this normalized function. So I have just called the function right now, I have not used the function, I have just stored the function here in my environment. And then what I am doing, I am using another data set called db churn 2 because original data set, I am not handling that, all the conversion and etc. we will do it in churn 2, so I am creating another data set db churn 2. And here I am normalizing total charges, I am normalizing monthly charges, the continuous variables, other than the continuous variables, all other things we have categorized.

So monthly charges I am normalizing and I am also normalizing tenure. So these are some of the things that we have normalized now. So if I have normalized these guys now, then the next step is basically also, also you will see that there are in my data set db churn 2 there are lots of factor data, where there is no internet service. Now no internet service means no only, there is no point on having another... So what I am doing is for 1 to 5 and 7 to 16.

(Refer Slide Time: 18:25)

The screenshot displays the R Studio environment. The script editor contains the following R code:

```
39 }
40
41 db_churn2<-data.frame(db_churn)
42 db_churn2$TotalCharges=normalize(db_churn2$TotalCharges)
43 db_churn2$MonthlyCharges=normalize(db_churn2$MonthlyCharges)
44 db_churn2$tenure=normalize(db_churn2$tenure)
45
46 for(i in c(1:5,7:16))db_churn2[,i]=ifelse(db_churn2[,i]=="No internet
47 for(i in c(1:5,7:16))db_churn2[,i]=ifelse(db_churn2[,i]=="No phone s
48 for(i in c(1:5,7:16))db_churn2[,i]=as.factor(db_churn2[,i])
49
50 en
```

The Environment pane on the right shows the following data objects:

- Global Environment
- Data
 - db_churn 7032 obs. of 21 variables
 - db_churn2 7032 obs. of 21 variables
 - n_NA 11 obs. of 1 variable
 - prop 2 obs. of 3 variables
- Functions
 - normalize function (x)

The Console pane shows the output of the following commands:

```
> db_churn2<-data.frame(db_churn)
> View(db_churn2)
> names(db_churn2)
 [1] "customerID"      "gender"          "SeniorCitizen"
 [4] "Partner"         "dependents"      "tenure"
 [7] "PhoneService"    "MultipleLines"   "InternetService"
[10] "OnlineSecurity"  "OnlineBackup"    "DeviceProtection"
[13] "TechSupport"     "StreamingTV"     "StreamingMovies"
[16] "Contract"        "PaperlessBilling" "PaymentMethod"
[19] "MonthlyCharges" "TotalCharges"    "Churn"
```

A histogram titled "Histogram of db_churn\$tenure" is visible in the bottom right corner, showing the frequency distribution of the tenure variable.

What is 1 to 5 in db churn 2, db_churn2, if I just want to know the column names, so I will just write names of db churn 2, I am getting 1 to 5, that means customer ID, gender, senior citizen, partner and dependents, these are categorical variable and 7 to 16 means these 2, these are also categorical variables. In these categorical variables, the moment no internet service is there, we change it to no. If it is no phone service, change it to no and then change them to factor.

So this is the changes that I am trying to do here. So all that no internets service, no phone service convert it to no. And then correspondingly you changed all the categorical variables to factor variable. So that is something that I have done here. Now my job is to predict so that is why I will be creating our training and testing data. But by chance, if you remembered, by chance you, if your job was to explain, you would have done only whole data only, no training data, no testing data.

(Refer Slide Time: 19:31)

The screenshot displays the RStudio environment. The script editor contains the following R code:

```
52 if (!require("caret")) install.packages("caret")
53 library("caret")
54 set.seed(7)
55 trainId = createDataPartition(db_churn2$churn,
56                               p=0.7, list=FALSE, times=1)
57
58 db_train = db_churn2[trainId,]
59 db_test = db_churn2[-trainId,]
60
61
62 #Decision Tree Model
```

The Environment pane on the right shows the following objects:

- Global Environment
- Data
 - db_churn: 7032 obs. of 21 variables
 - db_churn2: 7032 obs. of 21 variables
 - db_test: 2108 obs. of 21 variables
 - db_train: 4924 obs. of 21 variables
 - n_NA: 11 obs. of 1 variable
 - prop: 2 obs. of 3 variables

The Console pane shows the output of the code execution:

```
C:/Users/Dell/Desktop/Week9/Session 1/ >
[995,] 1427
[996,] 1428
[997,] 1429
[998,] 1430
[999,] 1432
[1000,] 1433
[ reached getOption("max.print") -- omitted 3924 rows ]
> db_train = db_churn2[trainId,]
> db_test = db_churn2[-trainId,]
>
```

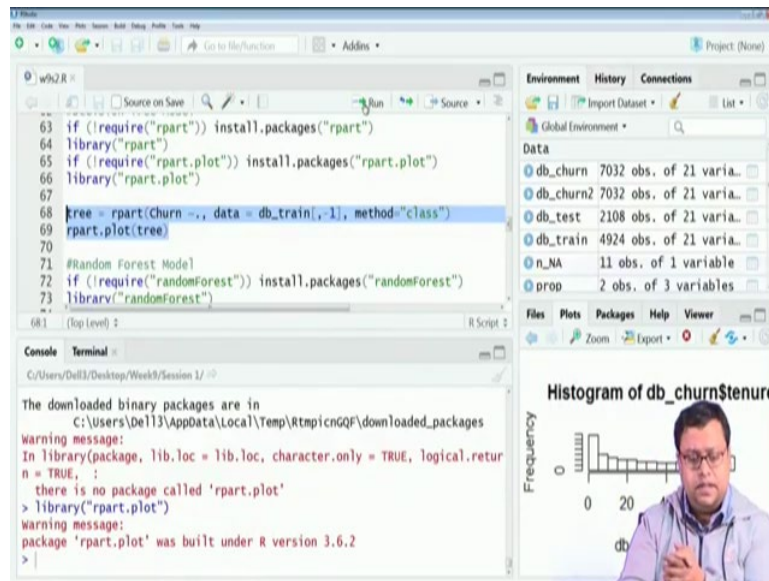
A histogram titled "Histogram of db_churn2\$tenure" is visible in the bottom right corner, showing the frequency distribution of the 'tenure' variable. A small inset video of a person is overlaid on the histogram.

So I will be requiring a library called caret, I am calling it, it is getting installed some, it will take 1 minute again. Yes, caret is getting installed and then I set seed 7, set seed 7 is so that it becomes reproducible research. Whatever result I am getting in my computer if I again run in my same computer, again once more, the same results will come. So that this randomness, randomly we will be creating training data, training testing, the randomness is fixed at a system level.

So then `trainId= createDataPartition(db_churn2$churn, p=0.7, list = FALSE, times=1)` , means with the 0.7 probability you put it into the training data. So I am creating training ID, using this training ID, training ID is nothing but the numbers which will be ultimately be chosen. So 3, 5, 6, 8, 9, some of the numbers in between has been dropped. And using these numbers, I am creating training data and testing data. So the numbers are going to training data and other numbers which are not here are going to testing data.

So, my training data has 4924 and testing data has 2108 observations. Now, I will create my model on the training data and I will predict it on the testing data.

(Refer Slide Time: 21:16)



The screenshot displays the R Studio environment. The script editor shows the following R code:

```
63 if (!require("rpart")) install.packages("rpart")
64 library("rpart")
65 if (!require("rpart.plot")) install.packages("rpart.plot")
66 library("rpart.plot")
67
68 tree = rpart(churn ~., data = db_train[,1], method="class")
69 rpart.plot(tree)
70
71 #Random Forest Model
72 if (!require("randomForest")) install.packages("randomForest")
73 library("randomForest")
```

The console shows the following output:

```
The downloaded binary packages are in
c:\Users\Dell13\AppData\Local\Temp\RtmpicnGQF\downloaded_packages
Warning message:
In library(package, lib.loc = lib.loc, character.only = TRUE, logical.return
n = TRUE, ...) :
there is no package called 'rpart.plot'
> library("rpart.plot")
Warning message:
package 'rpart.plot' was built under R version 3.6.2
>
```

The Environment pane on the right lists the following data objects:

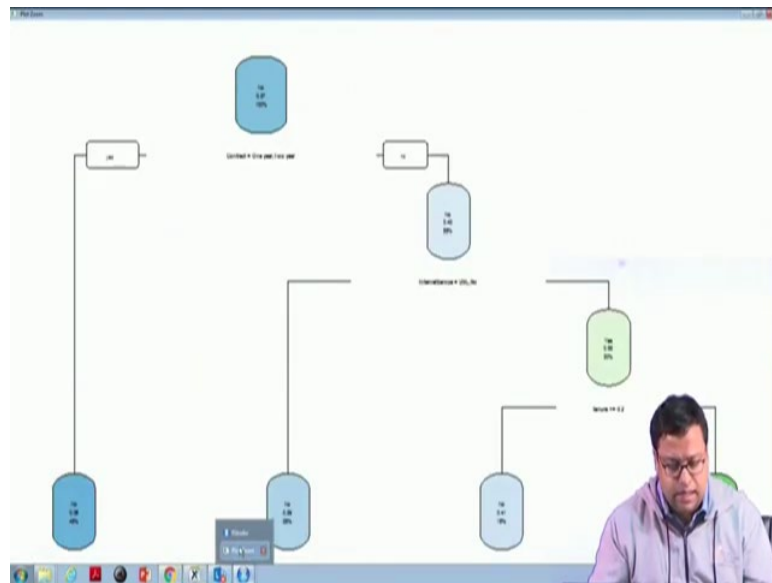
Object	Size
db_churn	7032 obs. of 21 variables
db_churn2	7032 obs. of 21 variables
db_test	2108 obs. of 21 variables
db_train	4924 obs. of 21 variables
n_NA	11 obs. of 1 variable
prop	2 obs. of 3 variables

A histogram titled "Histogram of db_churn\$tenure" is visible in the bottom right corner, showing the frequency distribution of the 'tenure' variable. A small inset image of a person is overlaid on the histogram.

Now, I here have chosen 3 models, one is decision tree random forest and logistic regression and testing I have done for all these 3 models. So, first is decision tree, so for that I require library rpart and rpart.plot. So, this is, these are the 2 libraries that has been downloaded. Now, I will run this dataset. So tree churn, so you in our data, I am not going into what exactly is random forest or what exactly is decision tree, I am not going into that, that should have been covered in a different course.

But this is actually creating a tree, where based on various kinds of rules you ultimately from the top node, you come to certain node and in that node based on the probability or based on the values that you have got, you find out the average of this particular node that is a the prediction. Whenever somebody comes in that node that particular value is assigned to him. So, that is called r decision tree. So, I will not go into the details of this entry, I will just run it and it is giving me a result like this.

(Refer Slide Time: 22:34)



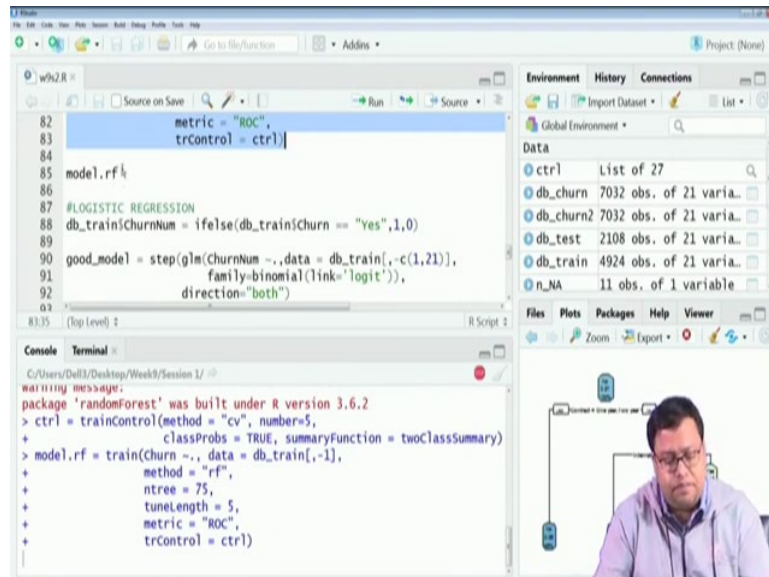
So, it is telling me that if contract 1 year or 2 year is no, if contract is 1 year or 2 year then if that is yes, so then you come here and otherwise you go there. And then you again check whether internet service is USL or no, if USL or no is yes, then you come here, if it is something else you come here and then if your tenure is greater than 0.2 I think then you come here, otherwise go there.

So, ultimate nodes are these 4 nodes this one, this one, this one and this one and here yes prediction is 0.7, here it is 0.41 here it is 0.29 and here is 0.6. So, the moment you are here, your probability of churning is very low. So now if you are have a contract of 1 year and 2 year, your chances of churning is very low. So the first job the company should do is creating a contract. By chance if he is not getting a contract then if he is using a data service, USL internet service, USL or no, there has chances of switching is low.

But by chance if he is using USL service, or is not using, so basically what is the other one, other than USL and no, what is the other one in internet service that is something that we have to also check.

giving you exact. So 0.204, that is the limit. So we will get the exact rules here. Now this is creating the tool only, that is all we are not predicting yet.

(Refer Slide Time: 25:28)



```
82 metric = "ROC",
83 trControl = ctrl]}
84
85 model.rf <-
86
87 #LOGISTIC REGRESSION
88 db_train$ChurnNum = ifelse(db_train$Churn == "Yes",1,0)
89
90 good_model = step(glm(ChurnNum ~.,data = db_train[,c(1,21)],
91 family=binomial(link="logit"),
92 direction="both")
93
94
95
96
```

Environment History Connections

Data

- ctrl List of 27
- db_churn 7032 obs. of 21 varia...
- db_churn2 7032 obs. of 21 varia...
- db_test 2108 obs. of 21 varia...
- db_train 4924 obs. of 21 varia...
- n_NA 11 obs. of 1 variable

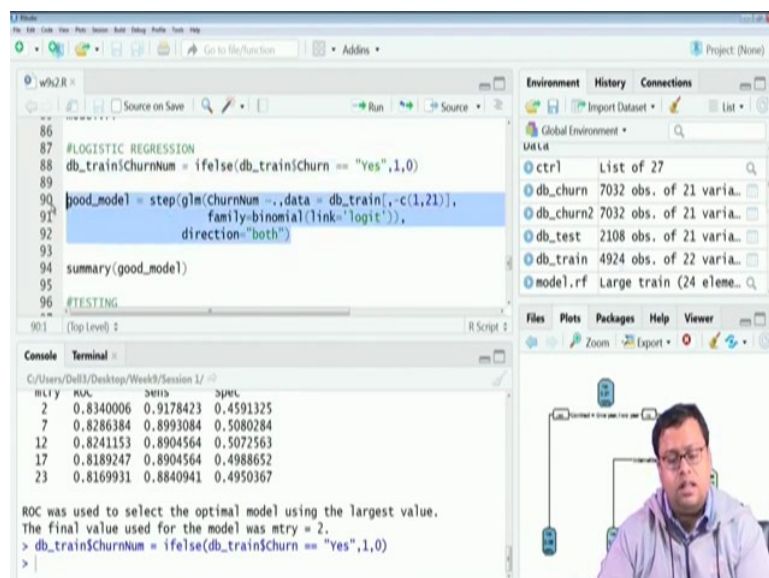
Files Plots Packages Help Viewer

Console Terminal

```
C:/Users/Dell3/Desktop/Week9/Session 1/ >
warning message:
package 'randomForest' was built under R version 3.6.2
> ctrl = trainControl(method = "cv", number=5,
+ classProbs = TRUE, summaryFunction = twoClassSummary)
> model.rf = train(churn ~., data = db_train[, -1],
+ method = "rf",
+ ntree = 75,
+ tuneLength = 5,
+ metric = "ROC",
+ trControl = ctrl)
```

The next is random forest, I will be using random forest. So random forest I have asked for this particular package to get downloaded and then I am using certain controls. And with those controls, I am running this training, where I am saying that I need 75 trees, tune length will be 5, metric will ROC to train the model. So it is training the model and the model specifications will be stored in model.rf. So that is something that will happen for some time, I will just wait. And it might not take much time because the data is not very big and it is done and we have the model details.

(Refer Slide Time: 26:09)



```
86
87 #LOGISTIC REGRESSION
88 db_train$ChurnNum = ifelse(db_train$Churn == "Yes",1,0)
89
90 good_model = step(glm(ChurnNum ~.,data = db_train[,c(1,21)],
91 family=binomial(link="logit"),
92 direction="both")
93
94 summary(good_model)
95
96 #TESTING
97
98
99
100
```

Environment History Connections

Data

- ctrl List of 27
- db_churn 7032 obs. of 21 varia...
- db_churn2 7032 obs. of 21 varia...
- db_test 2108 obs. of 21 varia...
- db_train 4924 obs. of 22 varia...
- model.rf Large train (24 eleme...

Files Plots Packages Help Viewer

Console Terminal

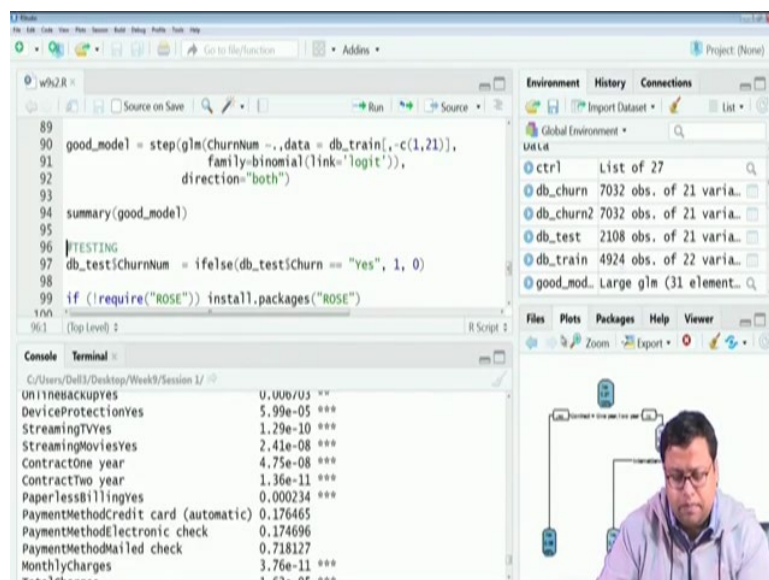
```
C:/Users/Dell3/Desktop/Week9/Session 1/ >
mtry | auc | dev |
2 | 0.8340006 | 0.9178423 | 0.4591325
7 | 0.8286384 | 0.8993084 | 0.5080284
12 | 0.8241153 | 0.8904564 | 0.5072563
17 | 0.8189247 | 0.8904564 | 0.4988652
23 | 0.8169931 | 0.8840941 | 0.4950367

ROC was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
> db_train$ChurnNum = ifelse(db_train$Churn == "Yes",1,0)
>
```

So it has been given that the sensitivity when the mtry is 2, 7, 12, 17, 23, how much is the sensitivity ROC values, that means that those things are given. And then our old school logistic regression. So I am breaking the data set between, I am changing the churn, remember, the churn till now was yes, no, I am making it 1, 0 so that I can use it in my logistic regression formula. 1, 0 and then I am using GLM and then I am also using a step within a GLM.

That means it will take all the variables in the first go and step function actually does the forward, actually backward movement. So it will drop 1 variable at a time to make sure that the aic value is maximized, so log likelihood value is maximized. So something like that.

(Refer Slide Time: 27:05)



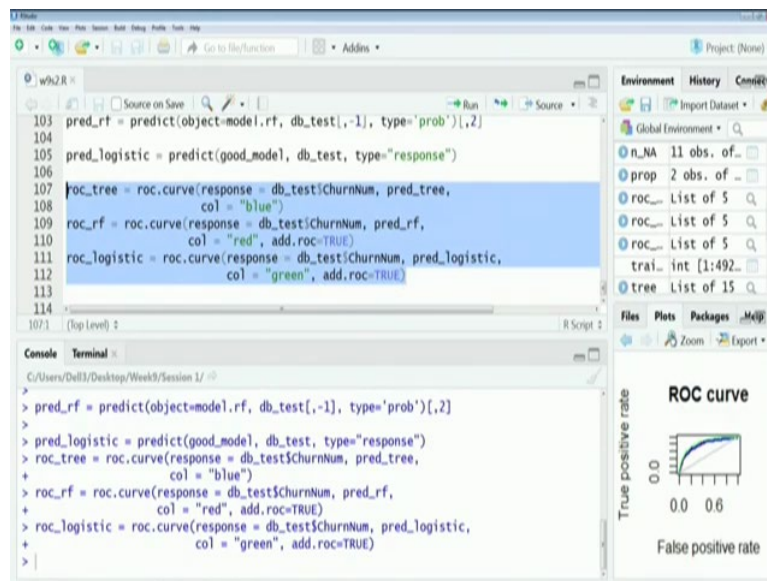
```
89
90 good_model = step(glm(ChurnNum ~., data = db_train[, -c(1,21)],
91                      family=binomial(link='logit'),
92                      direction="both")
93
94 summary(good_model)
95
96 #TESTING
97 db_test$ChurnNum = ifelse(db_test$Churn == "Yes", 1, 0)
98
99 if (!require("ROSE")) install.packages("ROSE")
100
```

Console Terminal

```
C:/Users/Dell/Desktop/Week9/Session 17 />
UnitBackups 0.006035 **
DeviceProtectionYes 5.99e-05 ***
StreamingTVYes 1.29e-10 ***
StreamingMoviesYes 2.41e-08 ***
ContractOne year 4.75e-08 ***
ContractTwo year 1.36e-11 ***
PaperlessBillingYes 0.000234 ***
PaymentMethodCredit card (automatic) 0.176465
PaymentMethodElectronic check 0.174696
PaymentMethodMailed check 0.718127
MonthlyCharges 3.76e-11 ***
ModelChance 1.63e-08 ***
```

So I will run, aic value should be minimized. So, okay I got it and the summary is something that I got here. So, some of the variables, actually most of the variables are significant and some if you are not significant. So I could have probably dropped payment method from the model to do it. Now, once this is done, once all the 3 models are done, I have to do the testing.

(Refer Slide Time: 27:30)

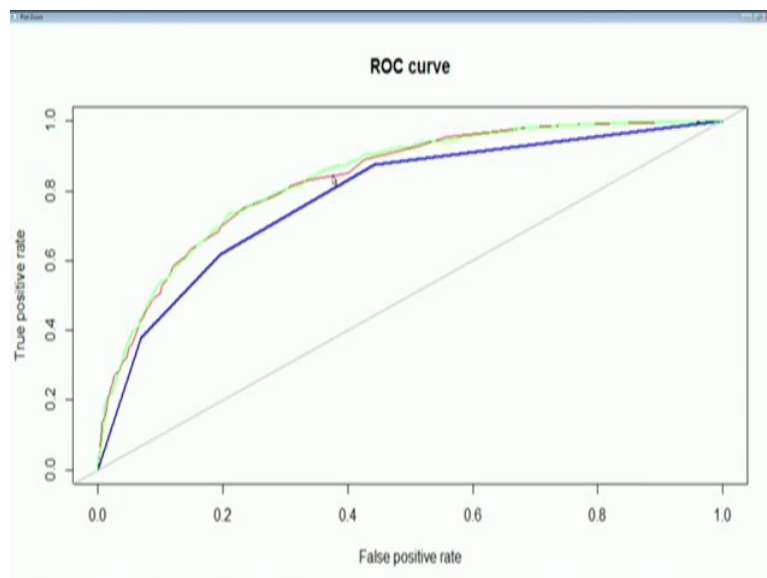


```
103 pred_rf = predict(object=model.rf, db_test[, -1], type='prob')[,2]
104
105 pred_logistic = predict(good_model, db_test, type="response")
106
107 roc_tree = roc.curve(response = db_test$ChurnNum, pred_tree,
108                      col = "blue")
109 roc_rf = roc.curve(response = db_test$ChurnNum, pred_rf,
110                   col = "red", add.roc=TRUE)
111 roc_logistic = roc.curve(response = db_test$ChurnNum, pred_logistic,
112                          col = "green", add.roc=TRUE)
113
114
1071 [top level]
```

The screenshot also shows a terminal window with the execution output and a small ROC curve plot in the bottom right corner.

So, to do that testing, I have to run this ROSE library and then predict one by one. So, I will run this ROSE library and for each of these 3 things predict 3, predict array, predict logistic I am predicting using the corresponding models. And once the prediction is done, I am plotting the ROC curves. Here if we just see that see the ROC curves, we will find out that what this mean? So, the first one I have taken blue, then red, then Green.

(Refer Slide Time: 28:01)



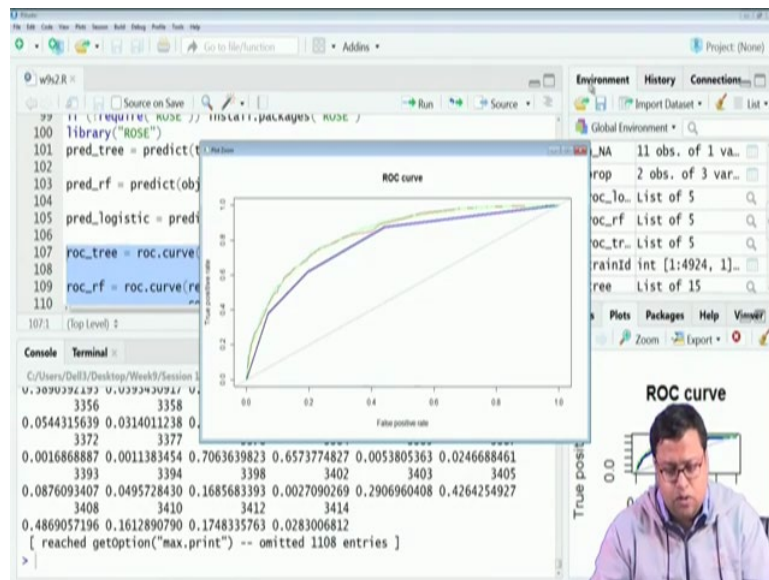
So if I just populate it, you will see that the blue line comes at the below, then the green line then red line and then green line, more or less red and green are same. Red and green are same means basically random forest and logistic function is same. Blue is the most, obviously

random forest is the ensemble model, it actually collects multiple trees data, and then does a prediction. So random forest will, very high probably will actually be better than tree.

But still, I can show you very simplistically, that logistic regression performs equally well. And in many practical situations that we use in our marketing, and that is why I focused on econometrics models and machine learning that the logistic regression or linear regression, basic regression techniques has higher explanation power and have higher prediction power also than certain machine learning techniques.

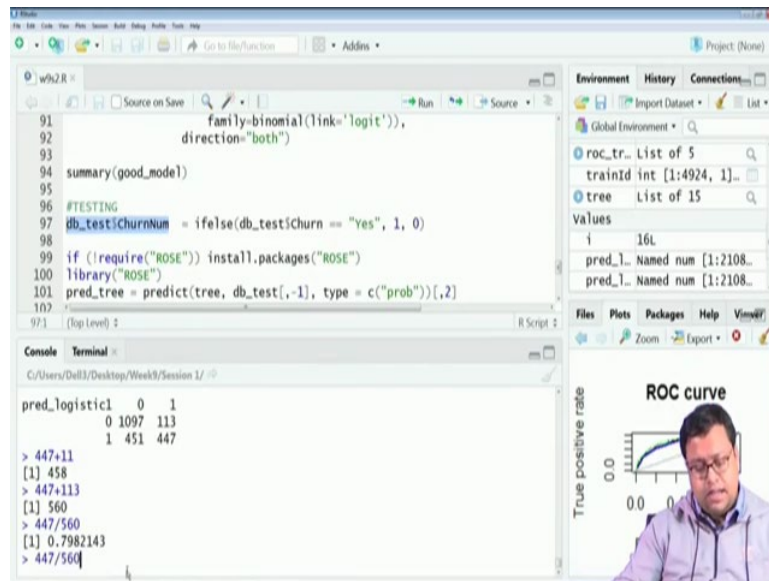
So, here also we can say that our thing is good enough. You could have drawn a confusion matrix also to check the case that how it is good, how it is bad.

(Refer Slide Time: 29:23)



So, to do that what I will do is you see that trade_logistic, this is basically a factor which is, which are basically probabilities. Now, if I just check that how much will I take, I can take anything between probably 0.3, 0.4 as my cutoff. If you remember in the earlier one, there was how much.

(Refer Slide Time: 29:55)



The screenshot shows the RStudio environment. The script editor contains the following R code:

```
91 family=binomial(link='logit'),
92 direction="both")
93
94 summary(good_model)
95
96 #TESTING
97 db_test$Churnum = ifelse(db_test$churn == "Yes", 1, 0)
98
99 if (!require("ROSE")) install.packages("ROSE")
100 library("ROSE")
101 pred_tree = predict(tree, db_test[, -1], type = c("prob"))[,2]
102
```

The Environment pane on the right shows the following objects:

- roc_tr_ List of 5
- trainId int [1:4924, 1]
- tree List of 15
- Values
 - i 16L
 - pred_1_ Named num [1:2108..
 - pred_1_ Named num [1:2108..

The Console pane shows the following output:

```
pred_logistic1 0 1
0 1097 113
1 451 447

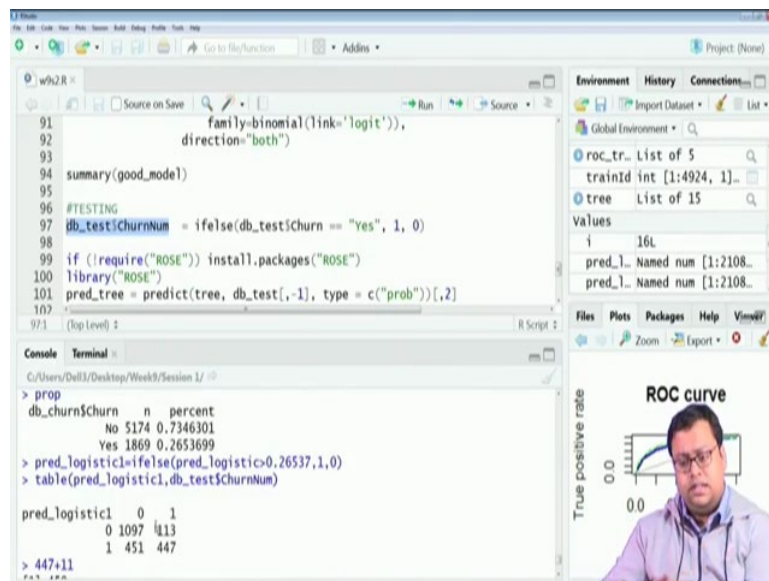
> 447+11
[1] 458
> 447+113
[1] 560
> 447/560
[1] 0.7982143
> 447/560
```

An ROC curve plot is visible in the bottom right corner, titled "ROC curve", showing True Positive Rate on the y-axis and False Positive Rate on the x-axis. A small inset image of a man is overlaid on the bottom right of the plot.

When we started, the analysis when we started, we had prop, we had around 0.26 that many yes and 0.73 is no. So maximum were no, a few were yes. So if I can just write that $\text{pred_logistic1} = \text{ifelse}(\text{pred_logistic} > 0.26537, 1, 0)$ And now if I just use this to create a crosstab, so I write $\text{table}(\text{pred_logistic}, \text{db_test}\$Churnum)$ and then run.

You will see that I am giving almost 447 times correct out of, so this is my predicted value, this is my actual value, so out of actually 113 and 447. So, $447 + 113$, which is 560, 447 divided by 560, these many times I am saying true positive. Out of all the positives, I am saying that allows, out of all the actual churns I am able to predict 79% of them using logistic regression alone.

(Refer Slide Time: 32:10)



The screenshot displays the RStudio environment. The script editor contains the following R code:

```
91 family=binomial(link='logit'),
92 direction="both")
93
94 summary(good_model)
95
96 #TESTING
97 db_test$ChurnNum = ifelse(db_test$Churn == "Yes", 1, 0)
98
99 if (!require("ROSE")) install.packages("ROSE")
100 library("ROSE")
101 pred_tree = predict(tree, db_test[, -1], type = c("prob"))[,2]
102
103
```

The console shows the following output:

```
C:/Users/Dell/Desktop/Week9/Session 1/ > prop
db_churn$Churn  n percent
No 5174 0.7346301
Yes 1869 0.2653699
> pred_logistic1=ifelse(pred_logistic>0.26537,1,0)
> table(pred_logistic1,db_test$ChurnNum)

pred_logistic1  0  1
0 1097  413
1  451 447
> 447+11
[1] 458
```

The Environment pane shows the following objects:

- roc_tr_ List of 5
- trainId int [1:4924, 1]
- tree List of 15

The Values pane shows:

- i 16L
- pred_1_ Named num [1:2108..
- pred_1_ Named num [1:2108..

The Console pane shows an ROC curve plot titled "ROC curve" with the y-axis labeled "True positive rate". A small inset image of a man is visible in the bottom right corner of the plot area.

So, these are some things that you have to check to see that whether you can do something or not. I can also use 0.5 as my prediction. So, in that case I have to check comma 0 and in this case it is much lesser, in this case it is much better. So 314 divided by 560, so I, should have taken the cutoff much lower and I have taken that and this is giving my optimal results.

So, optimal cutoff, you have to check the ROC curve but optimal curve is oftentimes what was there in the training data, how much percentage you have to divide, this probability and that percentage also. So, that is how we can predict, the explanation part is still not done yet, we will see in the rest of the part of this particular week, how I can explain the churning behavior or customer lifetime value and etc using marketing data. So, thank you for being with me. We will continue on this kind of topic in the next video.