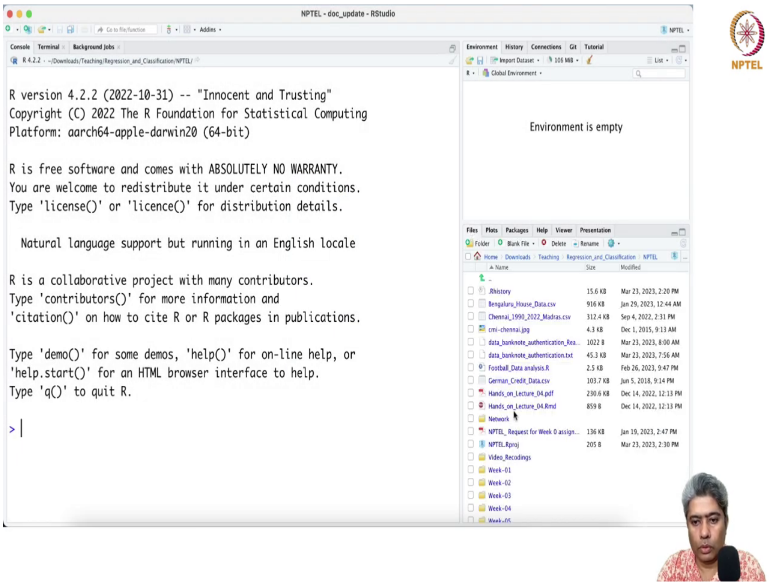
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 58
Hands on with R: Classify fake bank note with GLM

Hello all. Welcome to the part B of lecture 19. In this part we will do Hands on with R in this hands on video. We are going to do work on another data analysis where let me just first open the R.

(Refer Slide Time: 00:44)

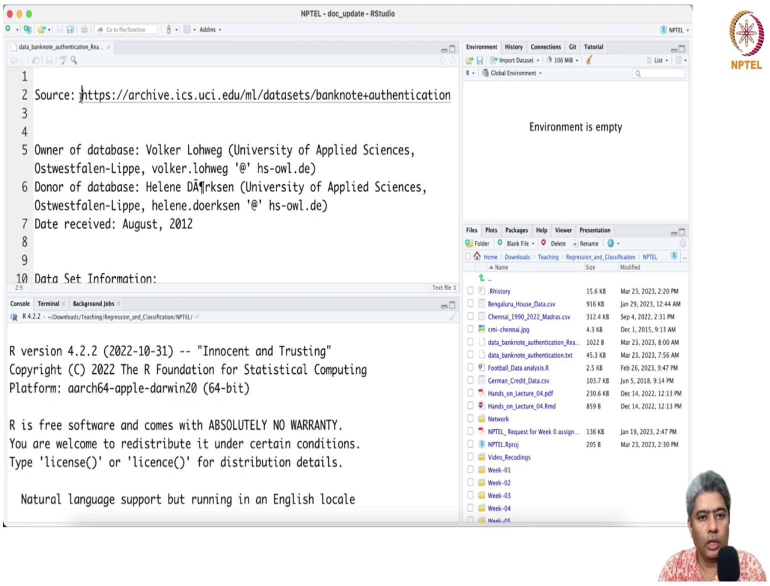


```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Environment is empty

Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

(Refer Slide Time: 00:58)



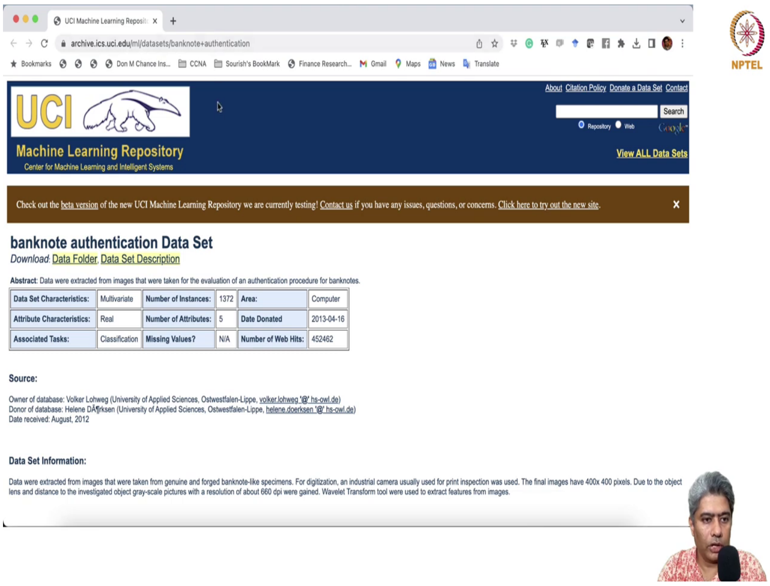
```
Source: https://archive.ics.uci.edu/ml/datasets/banknote+authentication


Owner of database: Volker Lohweg (University of Applied Sciences, Ostwestfalen-Lippe, volker.lohweg '@' hs-owl.de)
Donor of database: Helene DÃ¶rksen (University of Applied Sciences, Ostwestfalen-Lippe, helene.doerksen '@' hs-owl.de)
Date received: August, 2012


Data Set Information:
```

```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale
```

Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

So, what I am going to do here we have the from the UCI machine learning repository. If we just go there and let me just open my net, yeah ok.

(Refer Slide Time: 01:18)



Check out the beta version of the new UCI Machine Learning Repository we are currently testing! Contact us if you have any issues, questions, or concerns. Click here to try out the new site.

banknote authentication Data Set

Download: Data Folder, Data Set Description

Abstract: Data were extracted from images that were taken for the evaluation of an authentication procedure for banknotes.

Data Set Characteristics:	Multivariate	**Number of Instances:**	1372	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	5	**Date Donated**	2013-04-16
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	452462

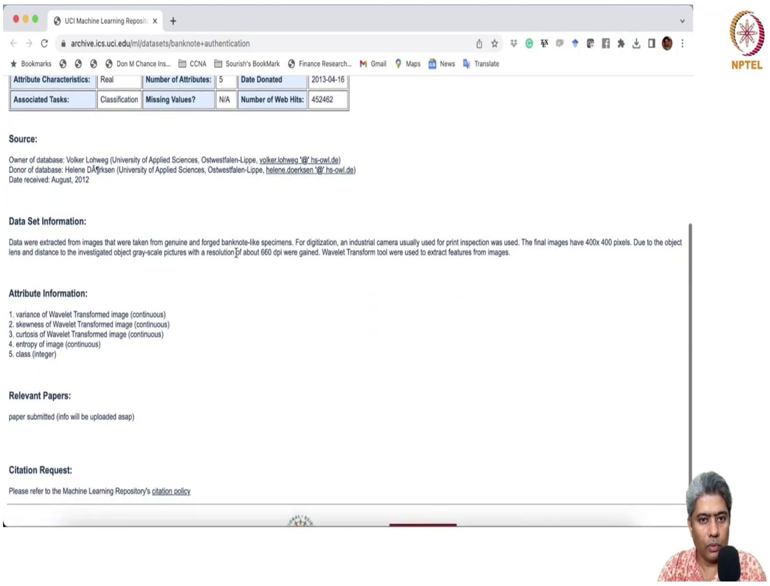
Source:

Owner of database: Volker Lohweg (University of Applied Sciences, Ostwestfalen-Lippe, volker.lohweg '@' hs-owl.de)
Donor of database: Helene DÃ¶rksen (University of Applied Sciences, Ostwestfalen-Lippe, helene.doerksen '@' hs-owl.de)
Date received: August, 2012

Data Set Information:

Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

(Refer Slide Time: 01:29)



Attribute Characteristics:	Real	Number of Attributes:	5	Date Donated	2013-04-16
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	452462

Source:

Owner of database: Volker Lohweg (University of Applied Sciences, Ostwestfalen-Lippe, volker.lohweg '@' hs-owl.de)
Donor of database: Helene DÃ¶rksen (University of Applied Sciences, Ostwestfalen-Lippe, helene.doerksen '@' hs-owl.de)
Date received: August, 2012

Data Set Information:

Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

Attribute Information:

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer)

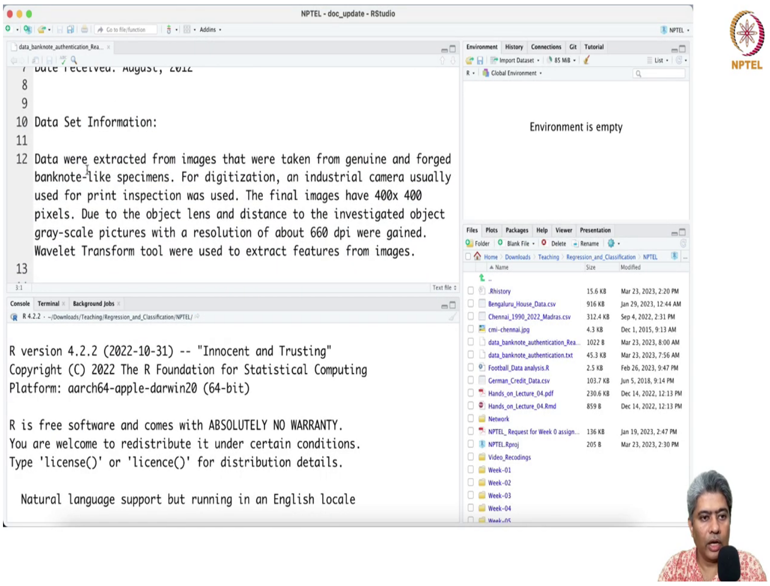
Relevant Papers:

paper submitted (info will be uploaded asap)

Citation Request:

Please refer to the Machine Learning Repository's citation policy

(Refer Slide Time: 01:34)



```
Date received: August, 2012


Data Set Information:

Data were extracted from images that were taken from genuine and forged
banknote-like specimens. For digitization, an industrial camera usually
used for print inspection was used. The final images have 400x 400
pixels. Due to the object lens and distance to the investigated object
gray-scale pictures with a resolution of about 660 dpi were gained.
Wavelet Transform tool were used to extract features from images.
```

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/

R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale
```

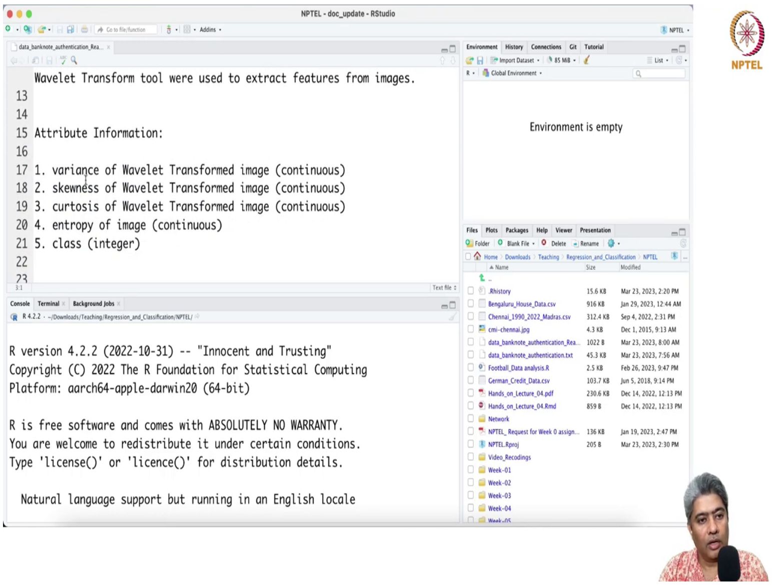
Home › Downloads › Teaching › Regression_and_Classification › NPTEL

Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

So, this is the UCI machine learning data repository. This is called bank note authentication data set. So, this is from Denmark. It says basically data were extracted from images that were taken from genuine and forged bank not like specimens. For digitization and industrial camera usually used for print inspection was used.

The final images have a 400 by 400 pixel due to object lens distance to the investigated object grayscale pictures with resolution of about 600 dpi where gained. Wavelet Transformation tool were used to extract the features from the images.

(Refer Slide Time: 02:15)



NPTEL - doc_update - RStudio
data_banknote_authentication_Rea...
Environment
History
Connections
Git
Tutorial
Import Dataset
Global Environment
Console
Terminal
Background Jobs
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
Files
Plots
Packages
Help
Viewer
Presentation
Folder
Blank File
Delete
Rename
Home › Downloads › Teaching › Regression_and_Classification › NPTEL
Text file

```
Wavelet Transform tool were used to extract features from images.


Attribute Information:

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer)
```

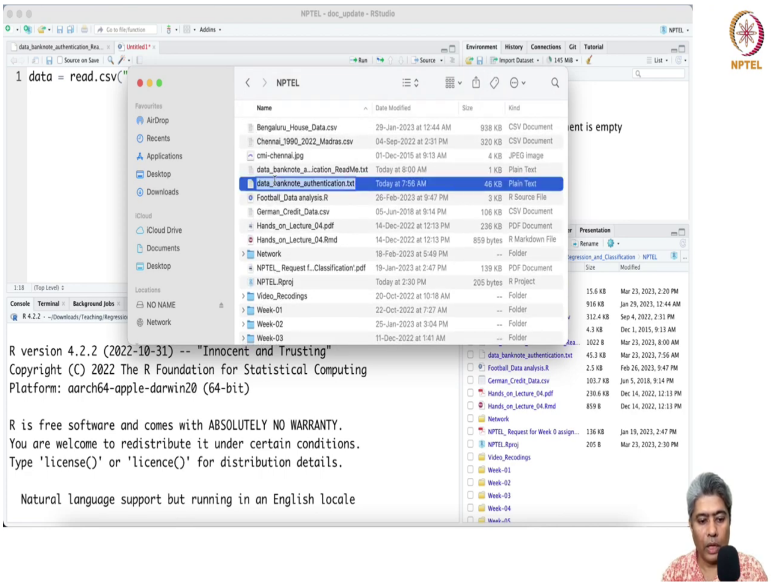
Environment is empty

```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale
```

Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

And these variants of wavelet transformations, skewness of the wavelet transformations, curtosis of the wavelet transformations, entropy of images and the classification manually done classification 0 or 1 that was done.

(Refer Slide Time: 02:40)



```
1 data = read.csv("
```

Name	Date Modified	Size	Kind
Bengaluru_House_Data.csv	29-Jan-2023 at 12:44 AM	938 KB	CSV Document
Chennai_1990_2022_Madras.csv	04-Sep-2022 at 2:31 PM	320 KB	CSV Document
cmi-chennai.jpg	01-Dec-2015 at 9:13 AM	4 KB	JPEG image
data_banknote_a...ication_ReadMe.txt	Today at 8:00 AM	1 KB	Plain Text
data_banknote_authentication.txt	Today at 7:56 AM	46 KB	Plain Text
Football_Data analysis.R	26-Feb-2023 at 9:47 PM	3 KB	R Source File
German_Credit_Data.csv	05-Jun-2018 at 9:14 PM	106 KB	CSV Document
Hands_on_Lecture_04.pdf	14-Dec-2022 at 12:13 PM	236 KB	PDF Document
Hands_on_Lecture_04.Rmd	14-Dec-2022 at 12:13 PM	859 bytes	R Markdown File
Network	18-Feb-2023 at 5:49 PM	--	Folder
NPTEL_ Request f...Classification'.pdf	19-Jan-2023 at 2:47 PM	139 KB	PDF Document
NPTEL.Rproj	Today at 2:30 PM	205 bytes	R Project
Video_Recodings	20-Oct-2022 at 10:18 AM	--	Folder
Week-01	22-Oct-2022 at 7:27 AM	--	Folder
Week-02	25-Jan-2023 at 3:04 PM	--	Folder
Week-03	11-Dec-2022 at 1:41 AM	--	Folder

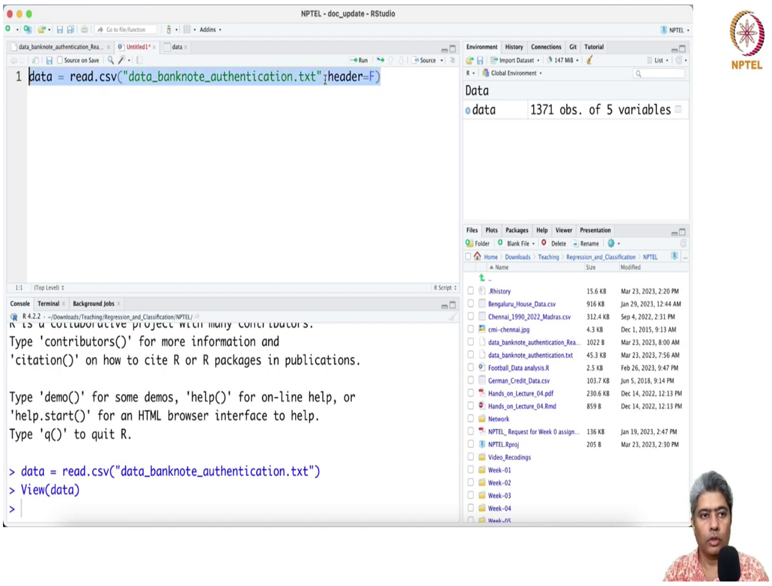
ent is empty

Name	Size	Modified
	15.6 KB	Mar 23, 2023, 2:20 PM
	916 KB	Jan 29, 2023, 12:44 AM
	312.4 KB	Sep 4, 2022, 2:31 PM
	4.3 KB	Dec 1, 2015, 9:13 AM
_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale
```

So, now let us start analysis. So, data equal to we will do read dot csv read dot csv. Let me just go quickly to the bank note and all downloads and classification. So, from so, this is the data set name. I have shared this data with in the swarm portal.

(Refer Slide Time: 03:25)



```
1 data = read.csv("data_banknote_authentication.txt",header=F)
```

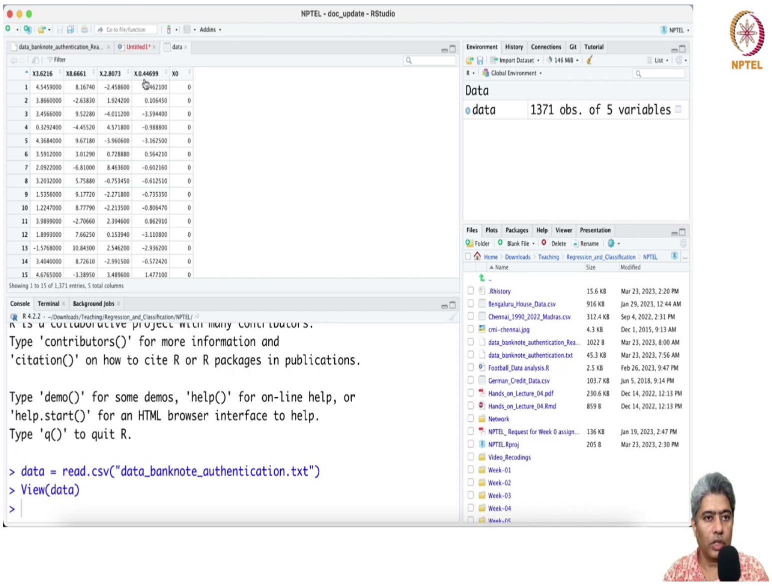
Data

data	1371 obs. of 5 variables

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data = read.csv("data_banknote_authentication.txt")
> View(data)
>
```

Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

(Refer Slide Time: 03:28)



	X3.6216	X8.6661	X.2.8073	X.0.44699	X0
1	4.5459000	8.16740	-2.458600	-1.462100	0
2	3.8660000	-2.63830	1.924200	0.106450	0
3	3.4566000	9.52280	-4.011200	-3.594400	0
4	0.3292400	-4.45520	4.571800	-0.988800	0
5	4.3684000	9.67180	-3.960600	-3.162500	0
6	3.5912000	3.01290	0.728880	0.564210	0
7	2.0922000	-6.81000	8.463600	-0.602160	0
8	3.2032000	5.75880	-0.753450	-0.612510	0
9	1.5356000	9.17720	-2.271800	-0.735350	0
10	1.2247000	8.77790	-2.213500	-0.806470	0
11	3.9899000	-2.70660	2.394600	0.862910	0
12	1.8993000	7.66250	0.153940	-3.110800	0
13	-1.5768000	10.84300	2.546200	-2.936200	0
14	3.4040000	8.72610	-2.991500	-0.572420	0
15	4.6765000	-3.38950	3.489600	1.477100	0

Showing 1 to 15 of 1,371 entries, 5 total columns

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data = read.csv("data_banknote_authentication.txt")
> View(data)
>
```

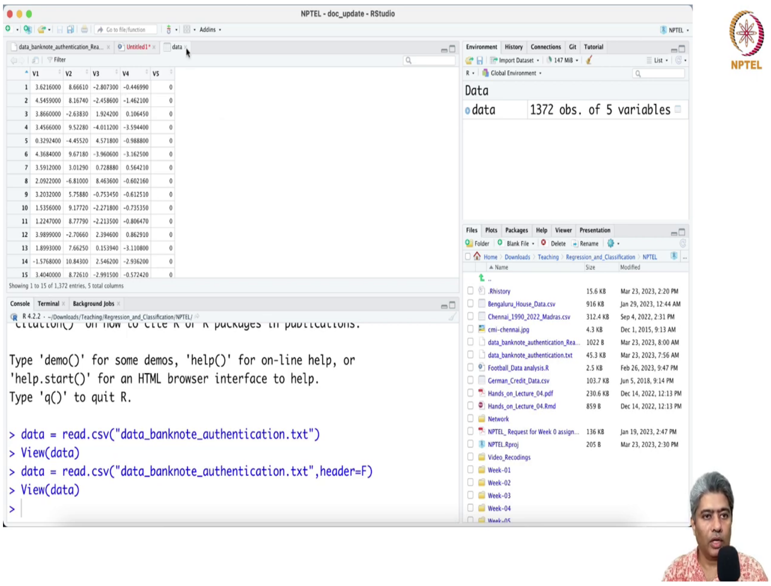
Data

data	1371 obs. of 5 variables

Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

So, if we just read this ok. So, it does not have any header because looks like this cannot be the name. So, it does not have any header.

(Refer Slide Time: 03:56)



	V1	V2	V3	V4	V5
1	3.6216000	8.66610	-2.807300	-0.446990	0
2	4.5459000	8.16740	-2.458600	-1.462100	0
3	3.8660000	-2.63830	1.924200	0.106450	0
4	3.4566000	9.52280	-4.011200	-3.594400	0
5	0.3292400	-4.45520	4.571800	-0.988800	0
6	4.3684000	9.67180	-3.960600	-3.162500	0
7	3.5912000	3.01290	0.728880	0.564210	0
8	2.0922000	-6.81000	8.463600	-0.602160	0
9	3.2032000	5.75880	-0.753450	-0.612510	0
10	1.5356000	9.17720	-2.271800	-0.735350	0
11	1.2247000	8.77790	-2.213500	-0.806470	0
12	3.9899000	-2.70660	2.394600	0.862910	0
13	1.8993000	7.66250	0.153940	-3.110800	0
14	-1.5768000	10.84300	2.546200	-2.936200	0
15	3.4040000	8.72610	-2.991500	-0.572420	0

Showing 1 to 15 of 1,372 entries, 5 total columns

Data

data 1372 obs. of 5 variables

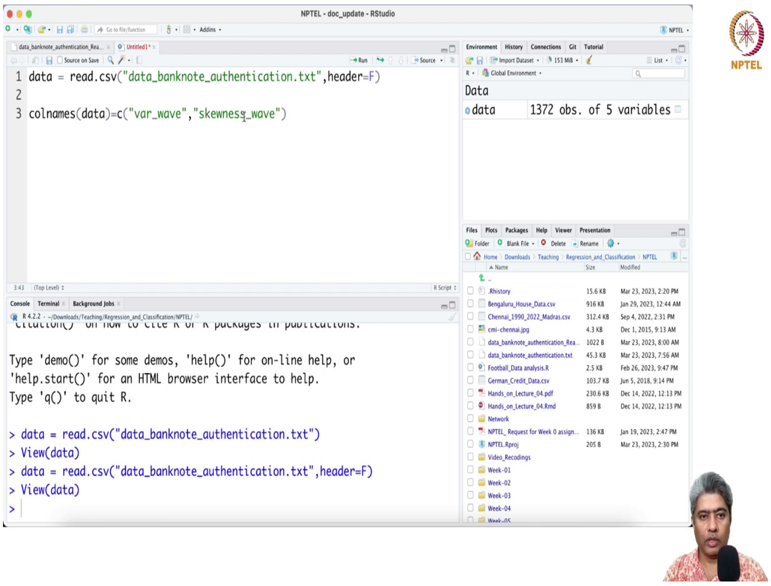
```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
 citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data = read.csv("data_banknote_authentication.txt")
> View(data)
> data = read.csv("data_banknote_authentication.txt",header=F)
> View(data)
>
```

Home › Downloads › Teaching › Regression_and_Classification › NPTEL

Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

So, we have to say header equal to header equal to false header equal to false and now we have the data correctly read, alright.

(Refer Slide Time: 04:04)



```r
data = read.csv("data_banknote_authentication.txt",header=F)

colnames(data)=c("var_wave","skewness_wave")
```

Data

data 1372 obs. of 5 variables

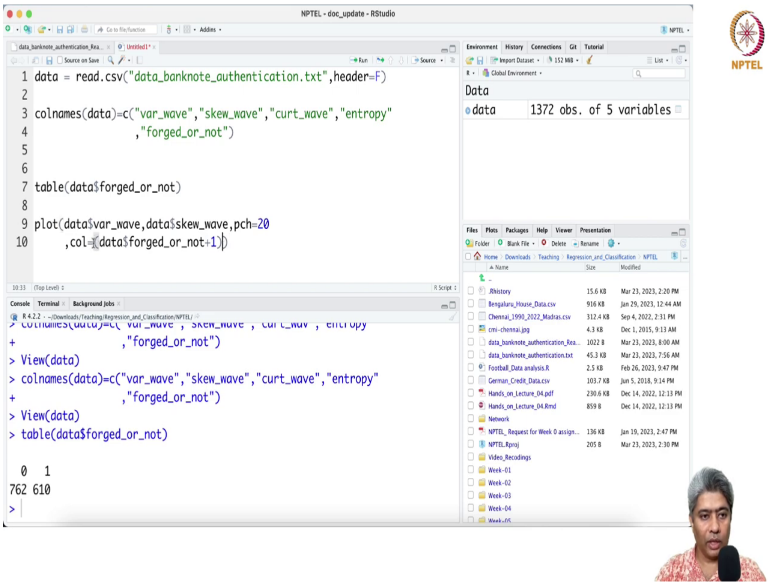
```
citation() on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> data = read.csv("data_banknote_authentication.txt")
> View(data)
> data = read.csv("data_banknote_authentication.txt",header=F)
> View(data)
>
```

Home › Downloads › Teaching › Regression_and_Classification › NPTEL

Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

Now, what we have to do? We have to put the colnames data equal to c. First, let us take these name variants of wavelet. So, first one will be var of wavelet, second will be the skewness, curtosis and entropy, ok.

(Refer Slide Time: 04:45)



```r
data = read.csv("data_banknote_authentication.txt",header=F)

colnames(data)=c("var_wave","skew_wave","curt_wave","entropy"
                 ,"forged_or_not")


table(data$forged_or_not)

plot(data$var_wave,data$skew_wave,pch=20
     ,col=(data$forged_or_not+1))
```

```
> colnames(data)=c("var_wave","skew_wave","curt_wav","entropy"
+                  ,"forged_or_not")
> View(data)
> colnames(data)=c("var_wave","skew_wave","curt_wave","entropy"
+                  ,"forged_or_not")
> View(data)
> table(data$forged_or_not)

  0   1 
762 610 
> 
```

Data

data	1372 obs. of 5 variables

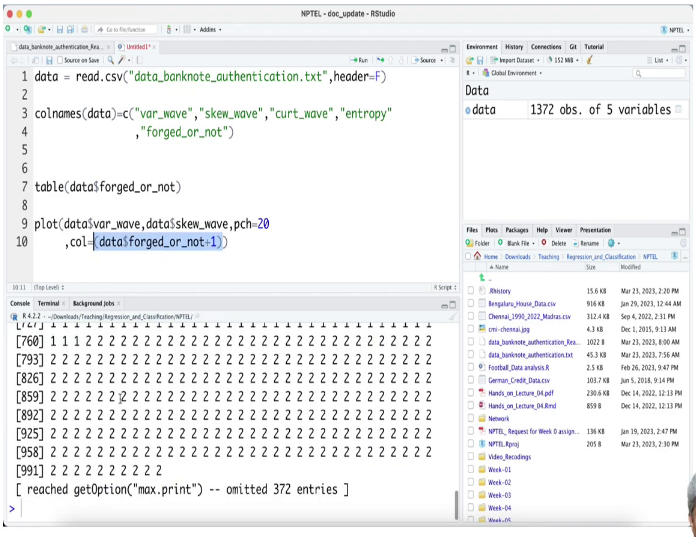
Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

Skewness of wavelet maybe I will just say skew. I do not want to put very large name. Skewness, curtosis of wavelets and “entropy”; and finally, last one is the class which is forged or not, right. So, let me just run this. So, now we have the these names are being given ok. So, maybe I will just say wave. Yeah. And now I think this is fine, right, ok.

Next, I will just have a look that how many forged cases and how many original note were there. So, out of 13,722 13,000, no, sorry, 1372 observations, 610 of them are forged and 762 of them are not forged. They are true original bank note. So, what I am going to do?

I am going to first plot data, dollar, variance of wave, comma, data, dollar skewness of wavelength maybe pch equal to 20. Now, what I am going to do is something you will I think you will have fun with it, data, dollar, you take the forge or not, see there is 0 or 1 is there. So, what I am going to do? I am going to add 1 to it, ok. I am going to add 1 to it.

(Refer Slide Time: 07:10)

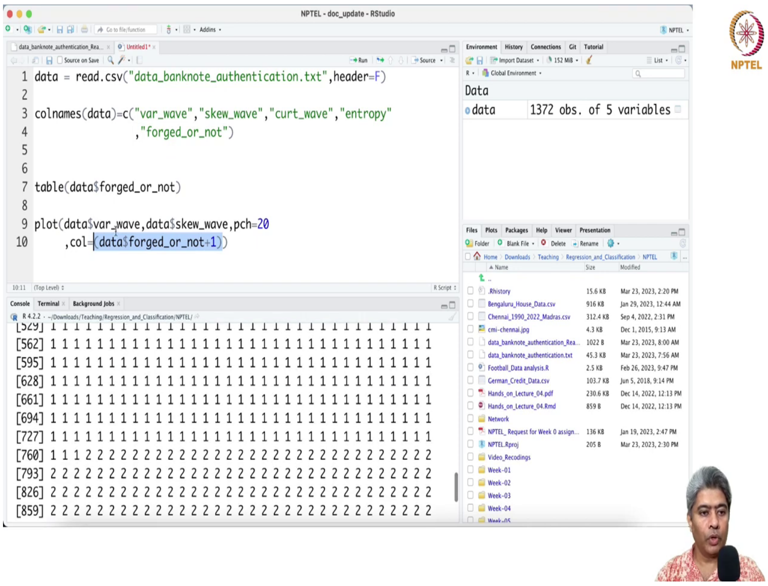


```
data = read.csv("data_banknote_authentication.txt",header=F)

colnames(data)=c("var_wave","skew_wave","curt_wave","entropy"
                 ,"forged_or_not")


table(data$forged_or_not)

plot(data$var_wave,data$skew_wave,pch=20
     ,col=(data$forged_or_not+1))
```

```
[760] 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[793] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[826] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[859] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[892] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[925] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[958] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[991] 2 2 2 2 2 2 2 2 2 2
[ reached getOption("max.print") -- omitted 372 entries ]
>
```

Data
data 1372 obs. of 5 variables

(Refer Slide Time: 07:15)



```r
data = read.csv("data_banknote_authentication.txt",header=F)

colnames(data)=c("var_wave","skew_wave","curt_wave","entropy"
                 ,"forged_or_not")



table(data$forged_or_not)

plot(data$var_wave,data$skew_wave,pch=20
     ,col=(data$forged_or_not+1))
```

```
[529] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[562] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[595] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[628] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[661] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[694] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[727] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[760] 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[793] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[826] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[859] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

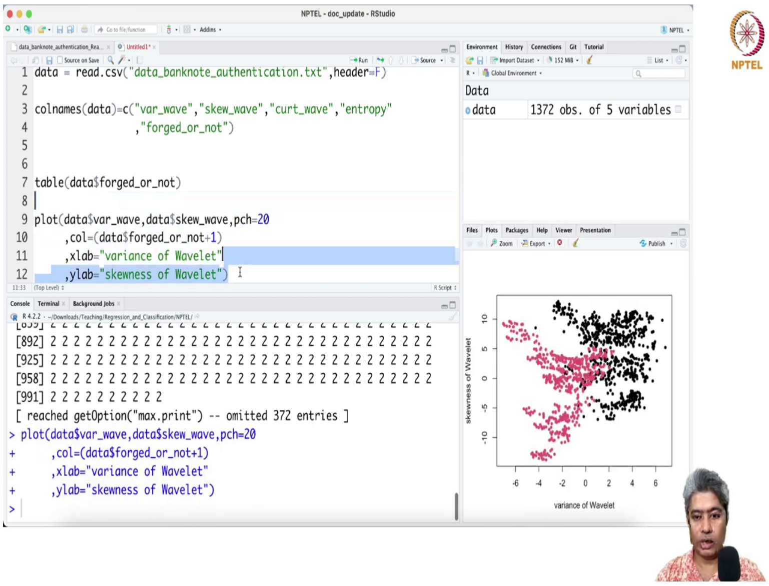
Data

data	1372 obs. of 5 variables

Name	Size	Modified
..		
.Rhistory	15.6 KB	Mar 23, 2023, 2:20 PM
Bengaluru_House_Data.csv	916 KB	Jan 29, 2023, 12:44 AM
Chennai_1990_2022_Madras.csv	312.4 KB	Sep 4, 2022, 2:31 PM
cmi-chennai.jpg	4.3 KB	Dec 1, 2015, 9:13 AM
data_banknote_authentication_Rea...	1022 B	Mar 23, 2023, 8:00 AM
data_banknote_authentication.txt	45.3 KB	Mar 23, 2023, 7:56 AM
Football_Data analysis.R	2.5 KB	Feb 26, 2023, 9:47 PM
German_Credit_Data.csv	103.7 KB	Jun 5, 2018, 9:14 PM
Hands_on_Lecture_04.pdf	230.6 KB	Dec 14, 2022, 12:13 PM
Hands_on_Lecture_04.Rmd	859 B	Dec 14, 2022, 12:13 PM
Network		
NPTEL_ Request for Week 0 assign...	136 KB	Jan 19, 2023, 2:47 PM
NPTEL.Rproj	205 B	Mar 23, 2023, 2:30 PM
Video_Recodings		
Week-01		
Week-02		
Week-03		
Week-04		
Week-05		

Now, what happens if I add 1 to it? What is happening? You actually is very simple. I am just adding either 1 or 0. It was 0 or 1. So, all 0s have become 1 and 1s have become 2. Now, as a color, 1 color equal to 1 means it is it will be black and color equal to 2 means it will be red, ok.

(Refer Slide Time: 07:30)



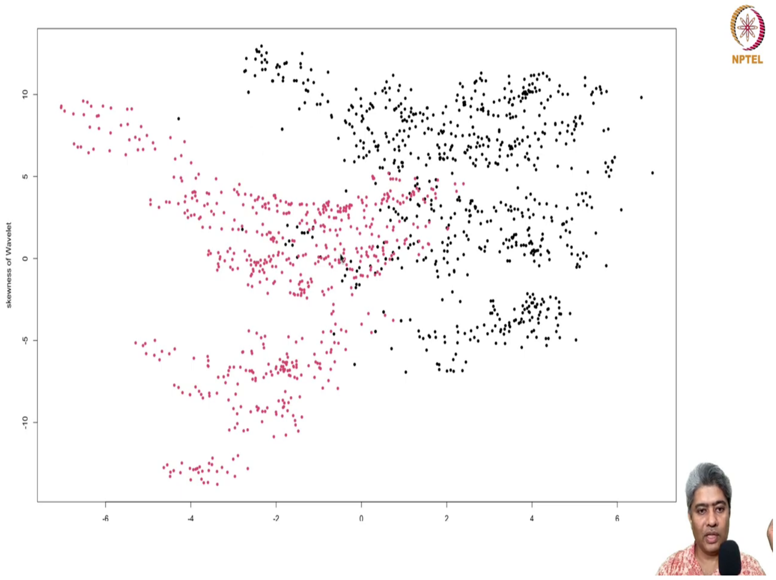
```r
data = read.csv("data_banknote_authentication.txt",header=F)

colnames(data)=c("var_wave","skew_wave","curt_wave","entropy"
                 ,"forged_or_not")


table(data$forged_or_not)

plot(data$var_wave,data$skew_wave,pch=20
     ,col=(data$forged_or_not+1)
     ,xlab="variance of Wavelet"
     ,ylab="skewness of Wavelet")
```

```
[892] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[925] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[958] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[991] 2 2 2 2 2 2 2 2 2 2
 [ reached getOption("max.print") -- omitted 372 entries ]
> plot(data$var_wave,data$skew_wave,pch=20
+      ,col=(data$forged_or_not+1)
+      ,xlab="variance of Wavelet"
+      ,ylab="skewness of Wavelet")
>
```

Data
data 1372 obs. of 5 variables

So, this is a this is I am taking making forged or not as the color that I am going to use xlab is let me just take simply “variance of Wavelet”, ok. And ylab, ylab equal to this is simply “skewness of Wavelet” transformed image, ok.

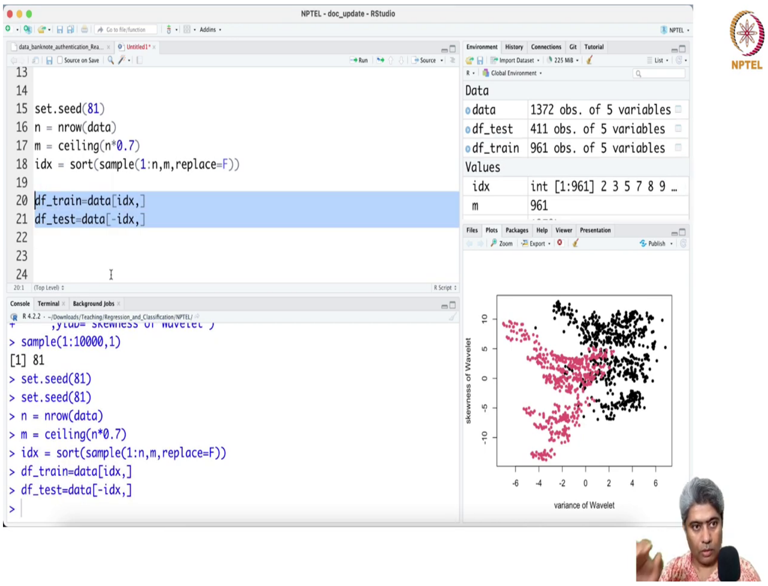
(Refer Slide Time: 08:06)



Let me just make this plot. Yeah, let me just zoom it. So, this is the data. So, now we see this data is very peculiar. The data has a very peculiar shape. You can see there is a very peculiar shape and there is bit of a overlap between the two kind of things. So, one is was forged, which was marked as red or these are like these are the specimens or points which have which represents forged note and the black ones are the one, which represent the original note.

So, there is we see there is a bit of a overlapping is happening, but mostly they are kind of well separated with slight overlapping between the 2. So, can we use wavelet variance and skewness of the wavelet these two feature to do the classification.

(Refer Slide Time: 09:19)



```
set.seed(81)
n = nrow(data)
m = ceiling(n*0.7)
idx = sort(sample(1:n,m,replace=F))

df_train=data[idx,]
df_test=data[-idx,]
```

Data

data	1372 obs. of 5 variables
df_test	411 obs. of 5 variables
df_train	961 obs. of 5 variables

Values

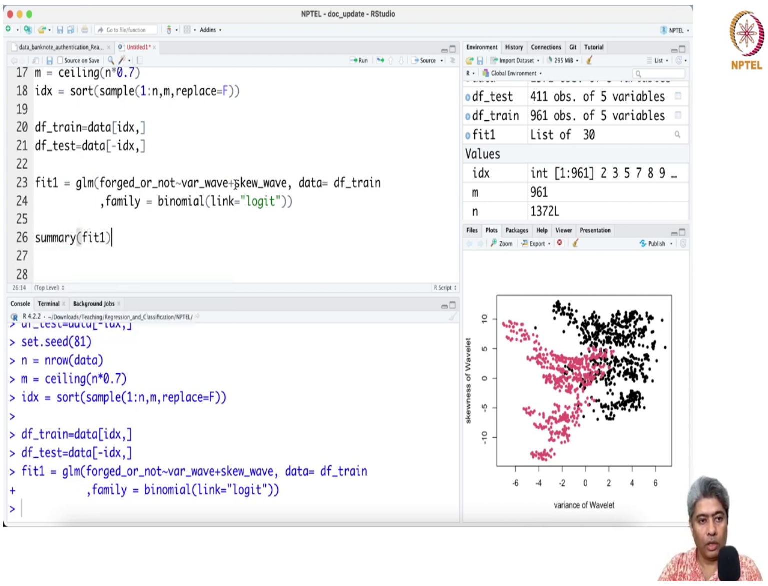
idx	int [1:961] 2 3 5 7 8 9 …
m	961

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
+       ,ylab= skewness of Wavelet )
> sample(1:10000,1)
[1] 81
> set.seed(81)
> set.seed(81)
> n = nrow(data)
> m = ceiling(n*0.7)
> idx = sort(sample(1:n,m,replace=F))
> df_train=data[idx,]
> df_test=data[-idx,]
>
```

So, what we will do we will first we will set a seed say sample first we need a set a seed. So, we need a number 1 is to say 1000. So, we will take 81. So, 81 as sitting our seed set dot seed ok, alright. And then n is the nrow of data and m is the ceiling of n times 0.7.

So, I am taking 70 percent for training data set and rest of them has test idx equal to sort, sample, 1, is to n comma m comma replace equal to false, ok. And then so, let me just have this. So, this idx is 961 cases have come and then now df train equal to data idx comma and df test equal to test equal to minus idx. So, let me just run the whole thing, yeah.

(Refer Slide Time: 11:10)



```
m = ceiling(n*0.7)
idx = sort(sample(1:n,m,replace=F))

df_train=data[idx,]
df_test=data[-idx,]

fit1 = glm(forged_or_not~var_wave+skew_wave, data= df_train
           ,family = binomial(link="logit"))

summary(fit1)
```

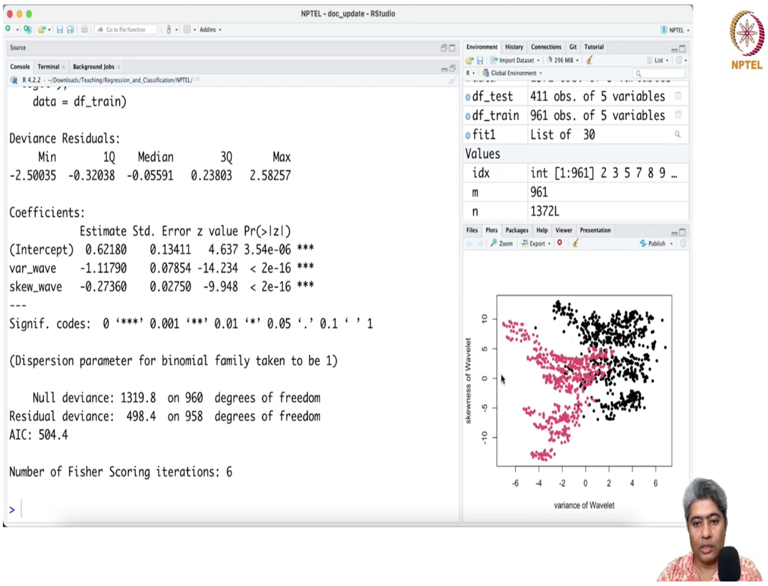
```
> df_test=data[-idx,]
> set.seed(81)
> n = nrow(data)
> m = ceiling(n*0.7)
> idx = sort(sample(1:n,m,replace=F))
>
> df_train=data[idx,]
> df_test=data[-idx,]
> fit1 = glm(forged_or_not~var_wave+skew_wave, data= df_train
+            ,family = binomial(link="logit"))
>
```

df_test	411 obs. of 5 variables
df_train	961 obs. of 5 variables
fit1	List of 30
Values	
idx	int [1:961] 2 3 5 7 8 9 …
m	961
n	1372L

So, first thing I am going to do I am going to fit a simple model with wavelet variance of wavelet and the skewness of wavelet as the you know feature of simple logistic regression fit. First model will going to be glm, ok glm. So, what we have is forged or not is as this comma variance of wavelet plus skewness of wavelet.

These two data equal to data equal to df train and then family equal to binomial link equal to "logit", alright. Now, this is what I am going to do. I am going to run this model. Now, if you run this summary fit1.

(Refer Slide Time: 12:34)

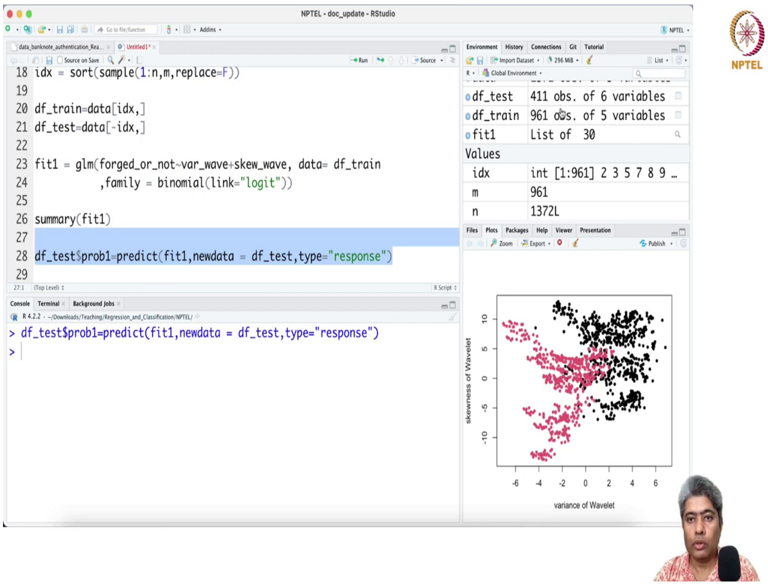


```
    data = df_train)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.50035  -0.32038  -0.05591   0.23803   2.58257  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.62180    0.13411   4.637 3.54e-06 ***
var_wave    -1.11790    0.07854 -14.234  < 2e-16 ***
skew_wave   -0.27360    0.02750  -9.948  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1319.8  on 960  degrees of freedom
Residual deviance:  498.4  on 958  degrees of freedom
AIC: 504.4

Number of Fisher Scoring iterations: 6

> 
```

```
df_test    411 obs. of 5 variables
df_train   961 obs. of 5 variables
fit1       List of  30
Values
  idx      int [1:961] 2 3 5 7 8 9 …
  m        961
  n        1372L
```

So, what you have it is saying that both variance of wavelet and the skewness of the wavelet is going to have a very strong accuracy sorry strong effect on the both variance of wavelet and the skewness of the wavelet is going to have a on the whether note will be forged or not, ok.

(Refer Slide Time: 13:08)



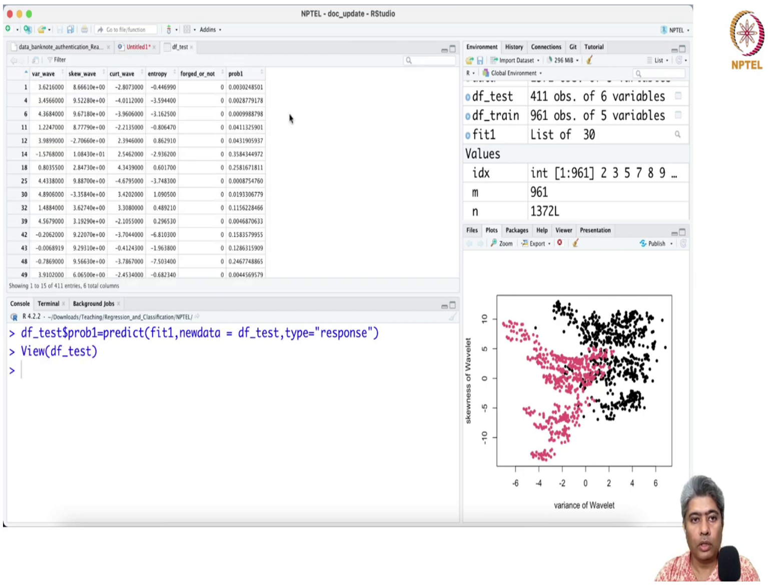
```
idx = sort(sample(1:n,m,replace=F))

df_train=data[idx,]
df_test=data[-idx,]

fit1 = glm(forged_or_not~var_wave+skew_wave, data= df_train
           ,family = binomial(link="logit"))

summary(fit1)

df_test$prob1=predict(fit1,newdata = df_test,type="response")

```

```
> df_test$prob1=predict(fit1,newdata = df_test,type="response")
>
```

df_test	411 obs. of 6 variables
df_train	961 obs. of 5 variables
fit1	List of 30
Values	
idx	int [1:961] 2 3 5 7 8 9 …
m	961
n	1372L

So, let us take that as and then now what we are going to do we are going to make the prediction on the test data set, ok. df test df test dollar probability1 equal to predict fit1, newdata equal to df test and then type equal to “response”, ok.

(Refer Slide Time: 13:54)



	var_wave	skew_wave	curt_wave	entropy	forged_or_not	prob1
1	3.6216000	8.66610e+00	-2.8073000	-0.446990	0	0.0030248501
4	3.4566000	9.52280e+00	-4.0112000	-3.594400	0	0.0028779178
6	4.3684000	9.67180e+00	-3.9606000	-3.162500	0	0.0009988798
11	1.2247000	8.77790e+00	-2.2135000	-0.806470	0	0.0411325901
12	3.9899000	-2.70660e+00	2.3946000	0.862910	0	0.0431905937
14	-1.5768000	1.08430e+01	2.5462000	-2.936200	0	0.3584344972
18	0.8035500	2.84730e+00	4.3439000	0.601700	0	0.2581671811
25	4.4338000	9.88700e+00	-4.6795000	-3.748300	0	0.0008754760
30	4.8906000	-3.35840e+00	3.4202000	1.090500	0	0.0193306779
32	1.4884000	3.62740e+00	3.3080000	0.489210	0	0.1156228466
39	4.5679000	3.19290e+00	-2.1055000	0.296530	0	0.0046870633
42	-0.2062000	9.22070e+00	-3.7044000	-6.810300	0	0.1583579955
43	-0.0068919	9.29310e+00	-0.4124300	-1.963800	0	0.1286315909
48	-0.7869000	9.56630e+00	-3.7867000	-7.503400	0	0.2467748865
49	3.9102000	6.06500e+00	-2.4534000	-0.682340	0	0.0044569579

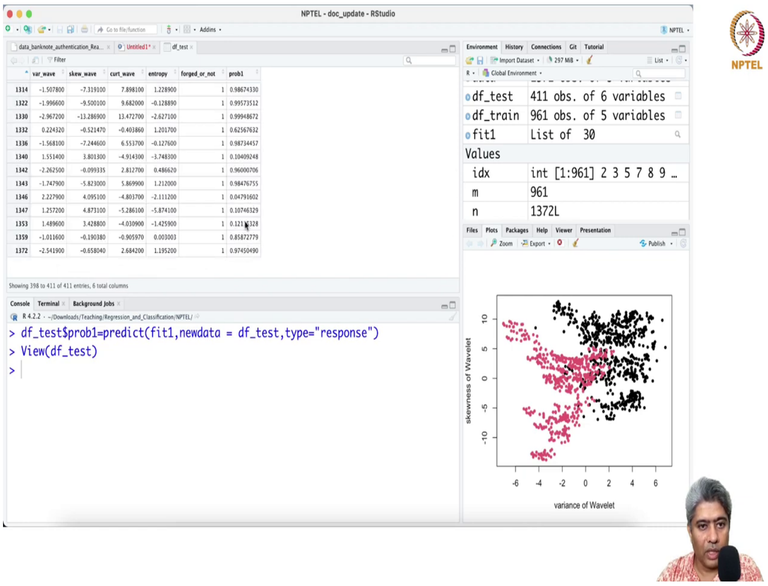
Showing 1 to 15 of 411 entries, 6 total columns

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
> df_test$prob1=predict(fit1,newdata = df_test,type="response")
> View(df_test)
> 
```

```
df_test    411 obs. of 6 variables
df_train   961 obs. of 5 variables
fit1       List of  30
Values
idx        int [1:961] 2 3 5 7 8 9 …
m          961
n          1372L
```

So, now if you see you see this, alright. So, now you can see that these values have come 0 and all this probability of 0.24.

(Refer Slide Time: 14:17)



	var_wave	skew_wave	curt_wave	entropy	forged_or_not	prob1
1314	-1.507800	-7.319100	7.898100	1.228900	1	0.98674330
1322	-1.996600	-9.500100	9.682000	-0.128890	1	0.99573512
1330	-2.967200	-13.286900	13.472700	-2.627100	1	0.99948672
1332	0.224320	-0.521470	-0.403860	1.201700	1	0.62567632
1336	-1.568100	-7.244600	6.553700	-0.127600	1	0.98734457
1340	1.551400	3.801300	-4.914300	-3.748300	1	0.10409248
1342	-2.262500	-0.099335	2.812700	0.486620	1	0.96000706
1343	-1.747900	-5.823000	5.869900	1.212000	1	0.98476755
1346	2.227900	4.095100	-4.803700	-2.111200	1	0.04791602
1347	1.257200	4.873100	-5.286100	-5.874100	1	0.10746329
1353	1.489600	3.428800	-4.030900	-1.425900	1	0.121[illegible]328
1359	-1.011600	-0.190380	-0.905970	0.003003	1	0.85872779
1372	-2.541900	-0.658040	2.684200	1.195200	1	0.97450490

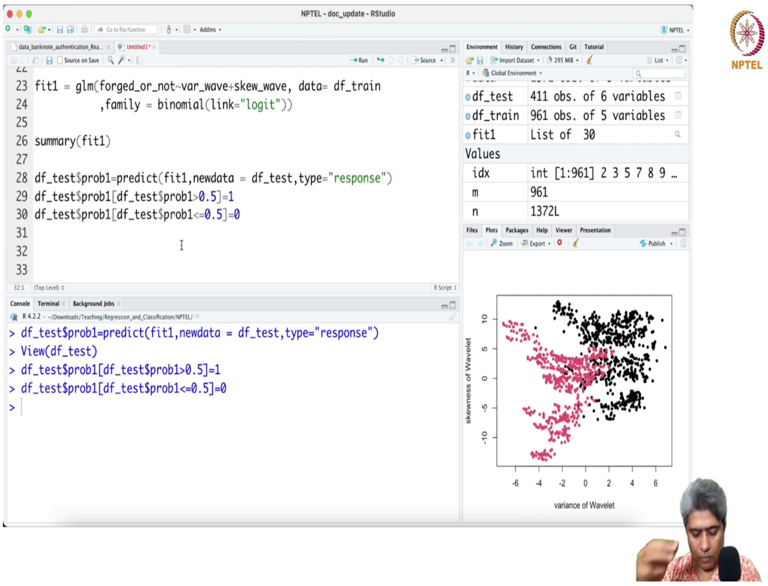
Showing 398 to 411 of 411 entries, 6 total columns

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
> df_test$prob1=predict(fit1,newdata = df_test,type="response")
> View(df_test)
>
```

```
df_test    411 obs. of 6 variables
df_train   961 obs. of 5 variables
fit1       List of  30
Values
 idx       int [1:961] 2 3 5 7 8 9 …
 m         961
 n         1372L
```

So, let us see if it is. So, mostly they are close to 0 some 0.1, 0.2 and if we just go down to 1 and you will see that they are reasonably close to 0.9, 0.8, 0.9, 0.8, right. Here is a one point where you got a you know 0.26, but the probability is less, but you know actually it is a forged note. So, there will be looks like there could be some misclassification possible. So, what we are going to do? We are going to make the prediction. These are the probability of this particular node to be forged note.

(Refer Slide Time: 14:51)



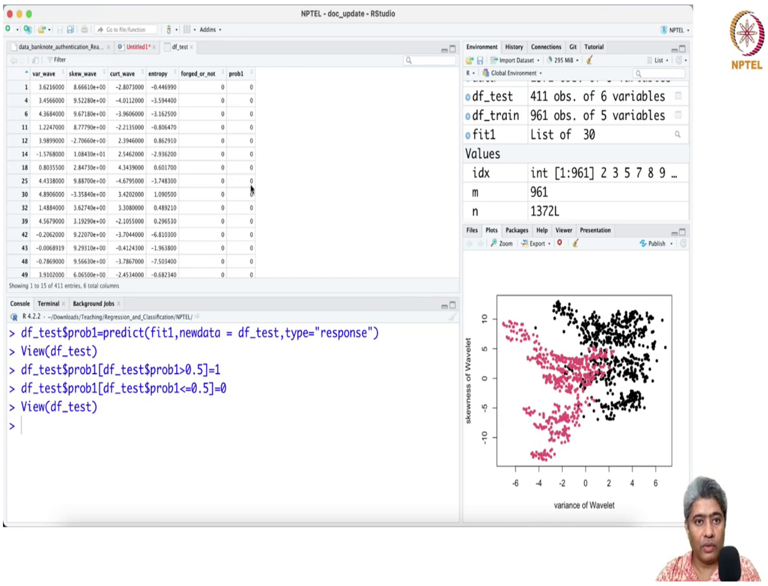
```
fit1 = glm(forged_or_not~var_wave+skew_wave, data= df_train
           ,family = binomial(link="logit"))

summary(fit1)

df_test$prob1=predict(fit1,newdata = df_test,type="response")
df_test$prob1[df_test$prob1>0.5]=1
df_test$prob1[df_test$prob1<=0.5]=0
```

df_test	411 obs. of 6 variables
df_train	961 obs. of 5 variables
fit1	List of 30
Values	
idx	int [1:961] 2 3 5 7 8 9 …
m	961
n	1372L

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
> df_test$prob1=predict(fit1,newdata = df_test,type="response")
> View(df_test)
> df_test$prob1[df_test$prob1>0.5]=1
> df_test$prob1[df_test$prob1<=0.5]=0
>
```

Now, I am going to make the prediction df testdollarprob1 equal to df testdollarprob1. If it is 0 greater than 0.5, we will call it 1. And if it is less than equal to 0.5, we will call it 0. So; that means, not forged ok, alright.

(Refer Slide Time: 15:32)



NPTEL - doc_update - RStudio

	var_wave	skew_wave	curt_wave	entropy	forged_or_not	prob1
1	3.6216000	8.66610e+00	-2.8073000	-0.446990	0	0
4	3.4566000	9.52280e+00	-4.0112000	-3.594400	0	0
6	4.3684000	9.67180e+00	-3.9606000	-3.162500	0	0
11	1.2247000	8.77790e+00	-2.2135000	-0.806470	0	0
12	3.9899000	-2.70660e+00	2.3946000	0.862910	0	0
14	-1.5768000	1.08430e+01	2.5462000	-2.936200	0	0
18	0.8035500	2.84730e+00	4.3439000	0.601700	0	0
25	4.4338000	9.88700e+00	-4.6795000	-3.748300	0	0
30	4.8906000	-3.35840e+00	3.4202000	1.090500	0	0
32	1.4884000	3.62740e+00	3.3080000	0.489210	0	0
39	4.5679000	3.19290e+00	-2.1055000	0.296530	0	0
42	-0.2062000	9.22070e+00	-3.7044000	-6.810300	0	0
43	-0.0068919	9.29310e+00	-0.4124300	-1.963800	0	0
48	-0.7869000	9.56630e+00	-3.7867000	-7.503400	0	0
49	3.9102000	6.06500e+00	-2.4534000	-0.682340	0	0

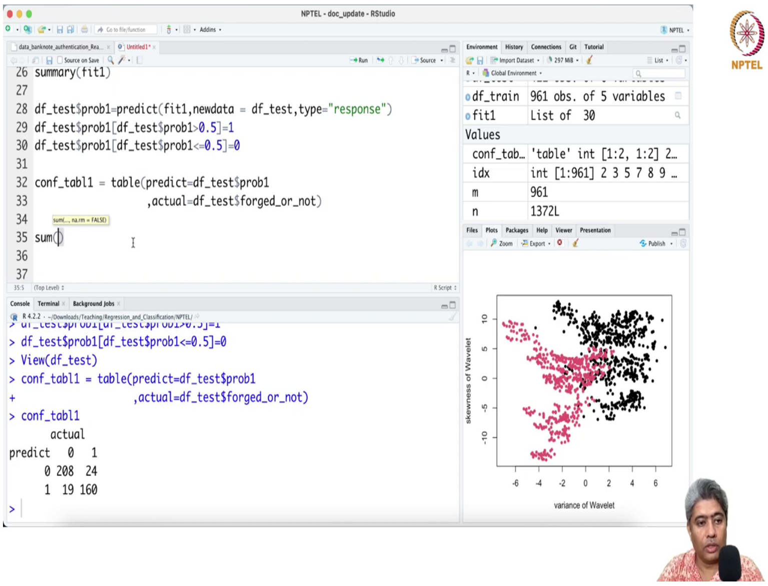
Showing 1 to 15 of 411 entries, 6 total columns

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
> df_test$prob1=predict(fit1,newdata = df_test,type="response")
> View(df_test)
> df_test$prob1[df_test$prob1>0.5]=1
> df_test$prob1[df_test$prob1<=0.5]=0
> View(df_test)
> 
```

df_test	411 obs. of 6 variables
df_train	961 obs. of 5 variables
fit1	List of 30
Values	
idx	int [1:961] 2 3 5 7 8 9 …
m	961
n	1372L

Now, what we are going to do? So, now, if I just look into the df test. So, now, they are all. So, here are there are some mis-classification we can see, but mostly they are agreeing.

(Refer Slide Time: 15:45)



```
summary(fit1)

df_test$prob1=predict(fit1,newdata = df_test,type="response")
df_test$prob1[df_test$prob1>0.5]=1
df_test$prob1[df_test$prob1<=0.5]=0

conf_tabl1 = table(predict=df_test$prob1
                   ,actual=df_test$forged_or_not)

sum()
```

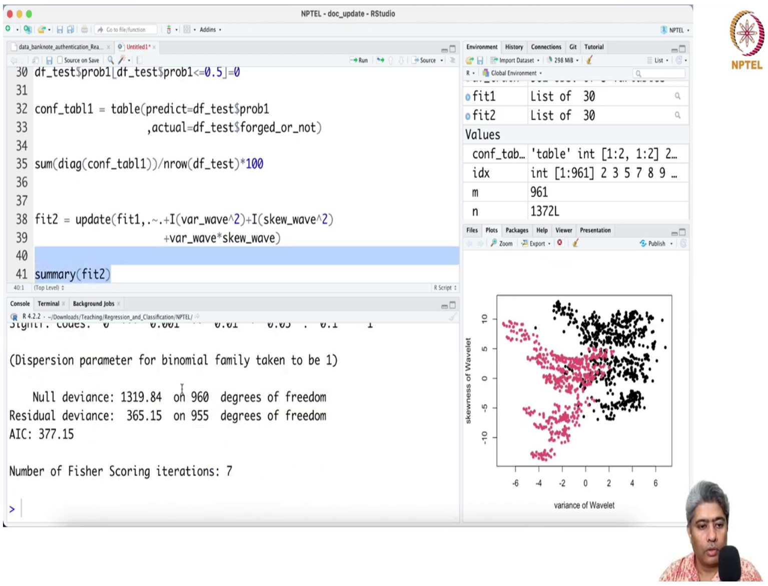
df_train	961 obs. of 5 variables
fit1	List of 30
Values	
conf_tab…	'table' int [1:2, 1:2] 2…
idx	int [1:961] 2 3 5 7 8 9 …
m	961
n	1372L

```
> df_test$prob1[df_test$prob1<=0.5]=0
> View(df_test)
> conf_tabl1 = table(predict=df_test$prob1
+                    ,actual=df_test$forged_or_not)
> conf_tabl1
       actual
predict   0   1
      0 208  24
      1  19 160
>
```

So, now we are going to I am going to calculate a confusion table, ok. So, table predict equal to df testdollarprob1 comma actual equal to df test dollar forged or not, ok. Now, if I just run the confusion table. So, this is my confusion table actually predicted 0 and actually 0 was 208 cases and in is the test data set. Similarly, predicted 1 and actually it was 1 was 160 cases.

There are 24 cases where it was predicted it is not forged, but it is actually forged and there are 19 cases where it was predicted forged, but it was not actually forged. So, there are possibility of and let us just check the accuracy.

(Refer Slide Time: 17:06)



```
df_test$prob1[df_test$prob1<=0.5]=0

conf_tabl1 = table(predict=df_test$prob1
                   ,actual=df_test$forged_or_not)

sum(diag(conf_tabl1))/nrow(df_test)*100


fit2 = update(fit1,.~.+I(var_wave^2)+I(skew_wave^2)
                     +var_wave*skew_wave)

summary(fit2)
```

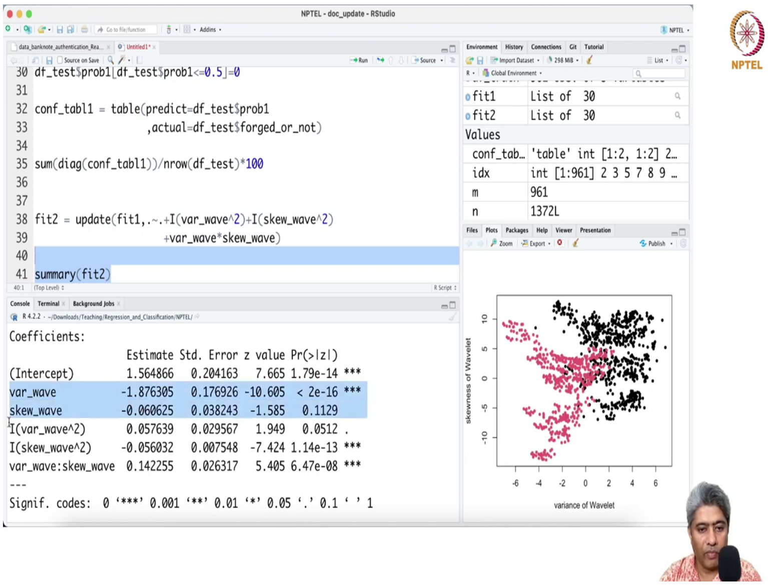
fit1	List of 30
fit2	List of 30
Values	
conf_tab…	'table' int [1:2, 1:2] 2…
idx	int [1:961] 2 3 5 7 8 9 …
m	961
n	1372L

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1319.84  on 960  degrees of freedom
Residual deviance:  365.15  on 955  degrees of freedom
AIC: 377.15

Number of Fisher Scoring iterations: 7

>
```

So, first I am going to take the diagonal of the confidence table. These are the correct predictions all correct prediction take the sum of that; sum of that. These are the all 368 are the total correct prediction and total number of all test is. Actually, just I can take nrow of df test is 411 of them there are 411 of them if I just make 100 out of this so, 89.53. So, this is my accuracy 89.53.

Now, if I make, I will make a second model fit2. What I will do is essentially I will just update fit 1 with same model plus few more engineered feature. So, I will just take variance of wave and square them plus a skewness of wave lengths and square them and finally, plus variance of wave times skewness of wave, ok. If I just run this and summary let me run the summary of fit2.

(Refer Slide Time: 19:20)



```
df_test$prob1[df_test$prob1<=0.5]=0

conf_tabl1 = table(predict=df_test$prob1
                   ,actual=df_test$forged_or_not)

sum(diag(conf_tabl1))/nrow(df_test)*100


fit2 = update(fit1,.~.+I(var_wave^2)+I(skew_wave^2)
                       +var_wave*skew_wave)

summary(fit2)
```

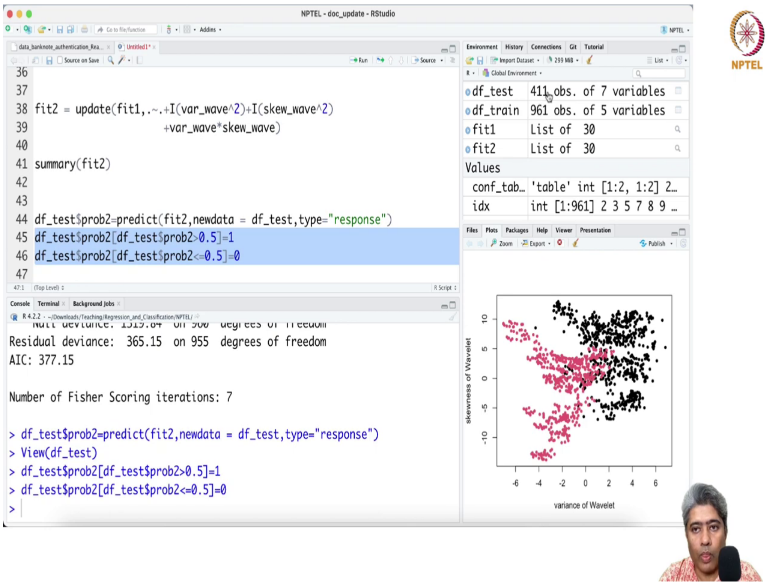
```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.564866   0.204163   7.665 1.79e-14 ***
var_wave          -1.876305   0.176926 -10.605  < 2e-16 ***
skew_wave         -0.060625   0.038243  -1.585   0.1129
I(var_wave^2)      0.057639   0.029567   1.949   0.0512 .
I(skew_wave^2)    -0.056032   0.007548  -7.424 1.14e-13 ***
var_wave:skew_wave 0.142255   0.026317   5.405 6.47e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

fit1	List of 30
fit2	List of 30
Values	
conf_tab…	'table' int [1:2, 1:2] 2…
idx	int [1:961] 2 3 5 7 8 9 …
m	961
n	1372L

So, looks like variance of wavelets and skewness of wavelet ok was. So, initially variance of wavelet ok if I just run the summary of fit1 we see that both wavelet both variance and skewness of the wavelet is as effect, but if I run on the forged or not whereas, if I run summary of fit2 where, it shows basically variance of wavelet does have a effect skewness does not, but the quadratic effect of skewness does have a effect and their interaction does have a effect.

So, this is very interesting phenomena, but let us see what is the whether adding this feature help me increasing the out of the sample accuracy or not. So, that will be my that will be the real test case.

(Refer Slide Time: 20:33)



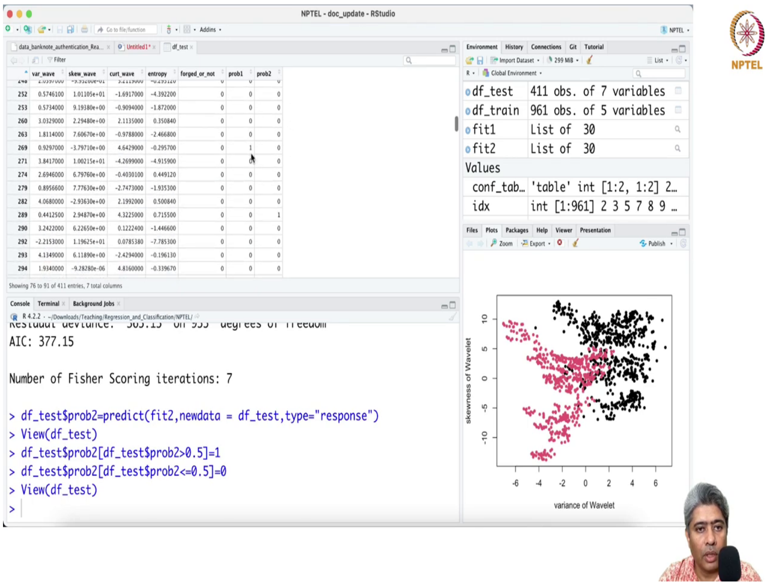
```
fit2 = update(fit1,.~.+I(var_wave^2)+I(skew_wave^2)
                      +var_wave*skew_wave)

summary(fit2)


df_test$prob2=predict(fit2,newdata = df_test,type="response")
df_test$prob2[df_test$prob2>0.5]=1
df_test$prob2[df_test$prob2<=0.5]=0
```

```
df_test    411 obs. of 7 variables
df_train   961 obs. of 5 variables
fit1       List of  30
fit2       List of  30
Values
conf_tab…  'table' int [1:2, 1:2] 2…
idx        int [1:961] 2 3 5 7 8 9 …
```

```
Residual deviance:  365.15  on 955  degrees of freedom
AIC: 377.15

Number of Fisher Scoring iterations: 7

> df_test$prob2=predict(fit2,newdata = df_test,type="response")
> View(df_test)
> df_test$prob2[df_test$prob2>0.5]=1
> df_test$prob2[df_test$prob2<=0.5]=0
>
```

So, what I will do? I will just take this guy copy and paste it here, but here instead of probability1 I have to take probability2 and then instead of fit2 I will just do fit1 I will just do fit2 it will add a new column it should add a new column the probability2, ok.

(Refer Slide Time: 20:47)


NPTEL - doc_update - RStudio
Go to file/function
Addins
data_banknote_authentication_Rea...
Untitled1*
df_test
Filter
NPTEL

	var_wave	skew_wave	curt_wave	entropy	forged_or_not	prob1	prob2
248	2.0397000	-9.95200e-01	3.2119000	-0.295120	0	0	0
252	0.5746100	1.01105e+01	-1.6917000	-4.392200	0	0	0
253	0.5734000	9.19380e+00	-0.9094000	-1.872000	0	0	0
260	3.0329000	2.29480e+00	2.1135000	0.350840	0	0	0
263	1.8114000	7.60670e+00	-0.9788000	-2.466800	0	0	0
269	0.9297000	-3.79710e+00	4.6429000	-0.295700	0	1	0
271	3.8417000	1.00215e+01	-4.2699000	-4.915900	0	0	0
274	2.6946000	6.79760e+00	-0.4030100	0.449120	0	0	0
279	0.8956600	7.77630e+00	-2.7473000	-1.935300	0	0	0
282	4.0680000	-2.93630e+00	2.1992000	0.500840	0	0	0
289	0.4412500	2.94870e+00	4.3225000	0.715500	0	0	1
290	3.2422000	6.22650e+00	0.1222400	-1.446600	0	0	0
292	-2.2153000	1.19625e+01	0.0785380	-7.785300	0	0	0
293	4.1349000	6.11890e+00	-2.4294000	-0.196130	0	0	0
294	1.9340000	-9.28280e-06	4.8160000	-0.339670	0	0	0

Showing 76 to 91 of 411 entries, 7 total columns

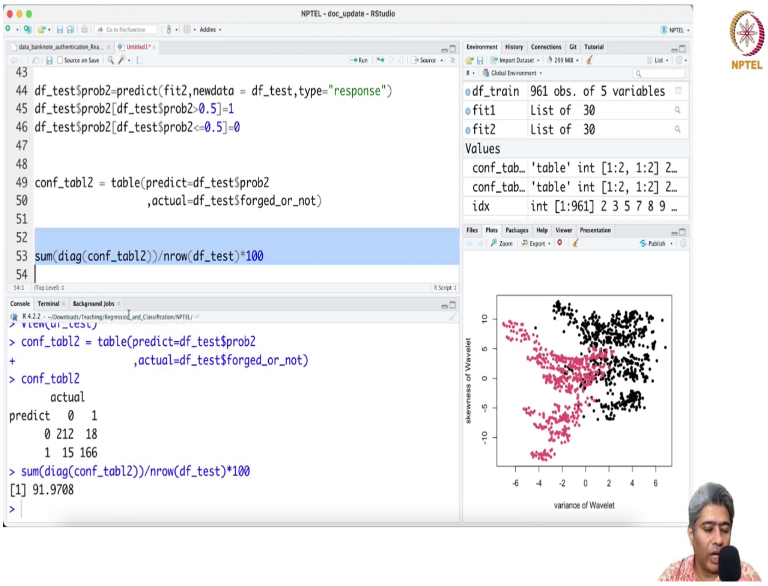
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/

```
Residual deviance:  363.15  on 955  degrees of freedom
AIC: 377.15

Number of Fisher Scoring iterations: 7

> df_test$prob2=predict(fit2,newdata = df_test,type="response")
> View(df_test)
> df_test$prob2[df_test$prob2>0.5]=1
> df_test$prob2[df_test$prob2<=0.5]=0
> View(df_test)
>
```

```
df_test     411 obs. of 7 variables
df_train    961 obs. of 5 variables
fit1        List of 30
fit2        List of 30
Values
conf_tab…   'table' int [1:2, 1:2] 2…
idx         int [1:961] 2 3 5 7 8 9 …
```

And then on the probability 2 I will make the prediction if it is more than 5, 0.5 I will say 1 otherwise I will say it is 0. So, now they are making 0 1, ok. Here is a misclassification; here was misclassification it was corrected here is a misclassification here is misclassification.

(Refer Slide Time: 21:26)



NPTEL - doc_update - RStudio

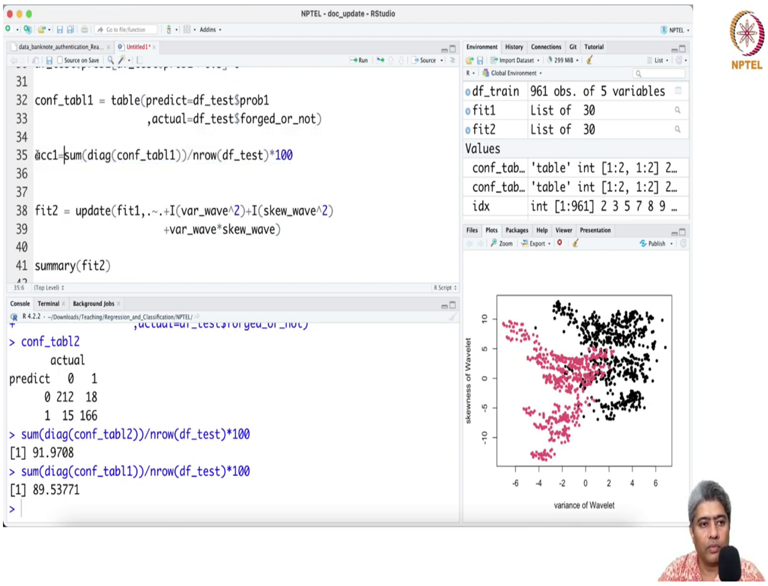
```r

df_test$prob2=predict(fit2,newdata = df_test,type="response")
df_test$prob2[df_test$prob2>0.5]=1
df_test$prob2[df_test$prob2<=0.5]=0


conf_tabl2 = table(predict=df_test$prob2
                   ,actual=df_test$forged_or_not)


sum(diag(conf_tabl2))/nrow(df_test)*100

```

```
> View(df_test)
> conf_tabl2 = table(predict=df_test$prob2
+                    ,actual=df_test$forged_or_not)
> conf_tabl2
       actual
predict   0   1
      0 212  18
      1  15 166
> sum(diag(conf_tabl2))/nrow(df_test)*100
[1] 91.9708
>
```

df_train	961 obs. of 5 variables
fit1	List of 30
fit2	List of 30
Values	
conf_tab…	'table' int [1:2, 1:2] 2…
conf_tab…	'table' int [1:2, 1:2] 2…
idx	int [1:961] 2 3 5 7 8 9 …

So, we will see and we can just calculate a confusion table simply second confusion table with probability2 and what is the situation of this is the second confusion table for this with the second model and if we just compute the accuracy simple accuracy this is 91.97. So, 92 percent accuracy we are getting whereas, for from this first model we are getting 89.5 percent.

(Refer Slide Time: 22:02)

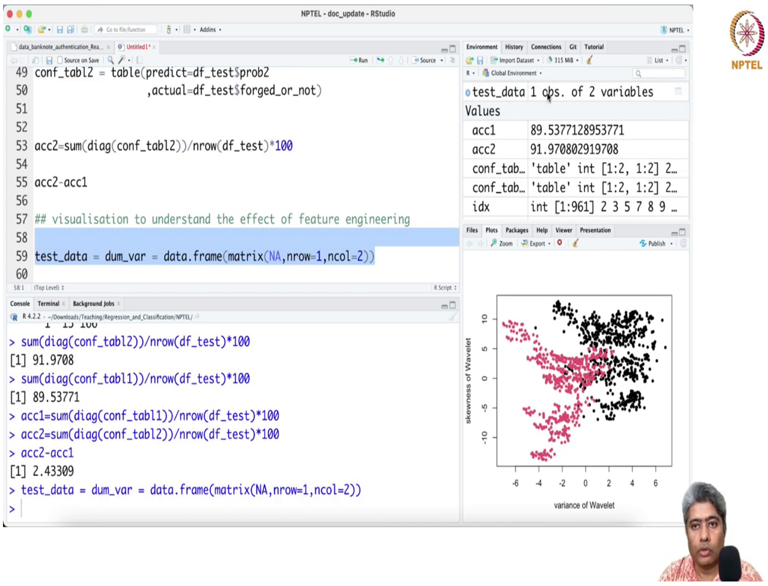


```
conf_tabl1 = table(predict=df_test$prob1
                   ,actual=df_test$forged_or_not)

acc1=sum(diag(conf_tabl1))/nrow(df_test)*100


fit2 = update(fit1,.~.+I(var_wave^2)+I(skew_wave^2)
                      +var_wave*skew_wave)

summary(fit2)
```

```
+                   ,actual=df_test$forged_or_not)
> conf_tabl2
       actual
predict   0   1
      0 212  18
      1  15 166
> sum(diag(conf_tabl2))/nrow(df_test)*100
[1] 91.9708
> sum(diag(conf_tabl1))/nrow(df_test)*100
[1] 89.53771
> 
```

```
df_train   961 obs. of 5 variables
fit1       List of  30
fit2       List of  30
Values
conf_tab…  'table' int [1:2, 1:2] 2…
conf_tab…  'table' int [1:2, 1:2] 2…
idx        int [1:961] 2 3 5 7 8 9 …
```

(Refer Slide Time: 22:23)



```r
conf_tabl2 = table(predict=df_test$prob2
                   ,actual=df_test$forged_or_not)


acc2=sum(diag(conf_tabl2))/nrow(df_test)*100

acc2-acc1

## visualisation to understand the effect of feature engineering

test_data = dum_var = data.frame(matrix(NA,nrow=1,ncol=2))

```

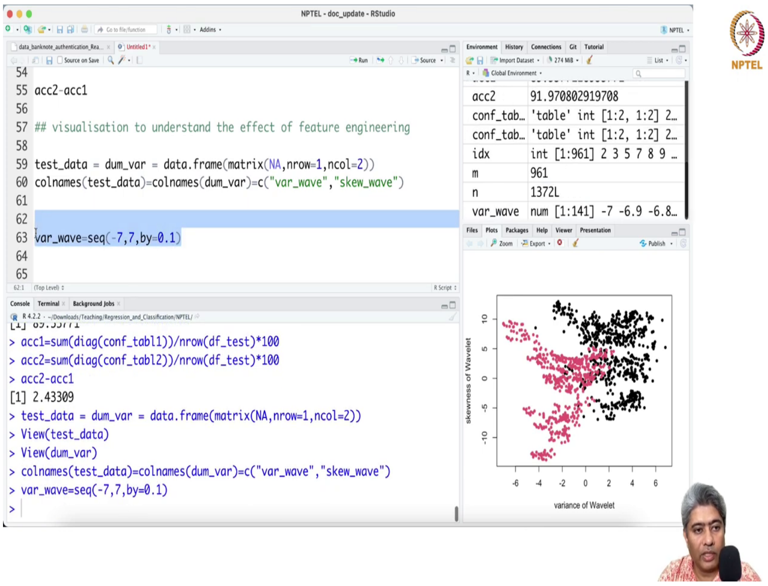
```
> sum(diag(conf_tabl2))/nrow(df_test)*100
[1] 91.9708
> sum(diag(conf_tabl1))/nrow(df_test)*100
[1] 89.53771
> acc1=sum(diag(conf_tabl1))/nrow(df_test)*100
> acc2=sum(diag(conf_tabl2))/nrow(df_test)*100
> acc2-acc1
[1] 2.43309
> test_data = dum_var = data.frame(matrix(NA,nrow=1,ncol=2))
>
```

```
test_data  1 obs. of 2 variables
Values
  acc1       89.5377128953771
  acc2       91.970802919708
  conf_tab…  'table' int [1:2, 1:2] 2…
  conf_tab…  'table' int [1:2, 1:2] 2…
  idx        int [1:961] 2 3 5 7 8 9 …
```

So, nearly 2.5 percent gain that we are getting, ok. So, this could be accuracy1 and this could be accuracy2 and accuracy2 minus accuracy1. So, 2.4 percent gain that we are getting because of the because of engineering proper feature engineering. Now, let us see how the geometry of the predictive space has changed using because of adding the feature engineer engineered feature.

So, I will do some visualisation to understand the effect of feature engineering, ok. Let us try to understand that. So, first I am going to do test data I am going to define a test data or dummy variable and dummy variable data dot frame and with that I will going to see matrix NA comma nrow equal to 1 ncol equal to 1 or sorry ncol equal to 2.

(Refer Slide Time: 23:50)



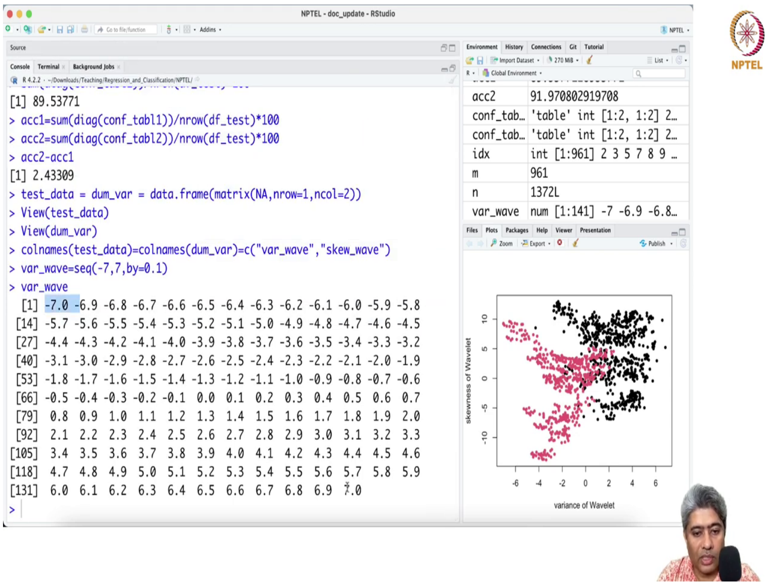
```
acc2-acc1

## visualisation to understand the effect of feature engineering

test_data = dum_var = data.frame(matrix(NA,nrow=1,ncol=2))
colnames(test_data)=colnames(dum_var)=c("var_wave","skew_wave")


var_wave=seq(-7,7,by=0.1)
```

```
acc2       91.970802919708
conf_tab…  'table' int [1:2, 1:2] 2…
conf_tab…  'table' int [1:2, 1:2] 2…
idx        int [1:961] 2 3 5 7 8 9 …
m          961
n          1372L
var_wave   num [1:141] -7 -6.9 -6.8…
```

```
> acc1=sum(diag(conf_tabl1))/nrow(df_test)*100
> acc2=sum(diag(conf_tabl2))/nrow(df_test)*100
> acc2-acc1
[1] 2.43309
> test_data = dum_var = data.frame(matrix(NA,nrow=1,ncol=2))
> View(test_data)
> View(dum_var)
> colnames(test_data)=colnames(dum_var)=c("var_wave","skew_wave")
> var_wave=seq(-7,7,by=0.1)
>
```

So, test data and this is the I am just creating a simply place holder, ok. Just a place holder and then I am going to add a colnames of a test data equal to colnames equal to dummy variable equals to this 2, ok. These are the then variance of wave, alright. So, variance now what I am going to do? So, variance of wavelet as a range of between see it is between minus 7 and 7 and skewness of wavelet is somewhere between maybe minus 12 to 12, ok. So, I am going to create this variance of wave define minus 7 to 7 by 0.1.

(Refer Slide Time: 25:12)

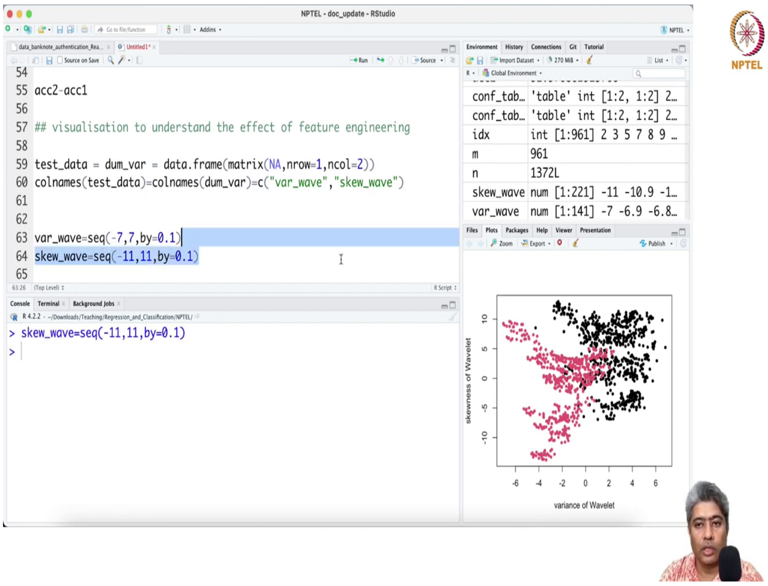


```
[1] 89.53771
> acc1=sum(diag(conf_tabl1))/nrow(df_test)*100
> acc2=sum(diag(conf_tabl2))/nrow(df_test)*100
> acc2-acc1
[1] 2.43309
> test_data = dum_var = data.frame(matrix(NA,nrow=1,ncol=2))
> View(test_data)
> View(dum_var)
> colnames(test_data)=colnames(dum_var)=c("var_wave","skew_wave")
> var_wave=seq(-7,7,by=0.1)
> var_wave
  [1] -7.0 -6.9 -6.8 -6.7 -6.6 -6.5 -6.4 -6.3 -6.2 -6.1 -6.0 -5.9 -5.8
 [14] -5.7 -5.6 -5.5 -5.4 -5.3 -5.2 -5.1 -5.0 -4.9 -4.8 -4.7 -4.6 -4.5
 [27] -4.4 -4.3 -4.2 -4.1 -4.0 -3.9 -3.8 -3.7 -3.6 -3.5 -3.4 -3.3 -3.2
 [40] -3.1 -3.0 -2.9 -2.8 -2.7 -2.6 -2.5 -2.4 -2.3 -2.2 -2.1 -2.0 -1.9
 [53] -1.8 -1.7 -1.6 -1.5 -1.4 -1.3 -1.2 -1.1 -1.0 -0.9 -0.8 -0.7 -0.6
 [66] -0.5 -0.4 -0.3 -0.2 -0.1  0.0  0.1  0.2  0.3  0.4  0.5  0.6  0.7
 [79]  0.8  0.9  1.0  1.1  1.2  1.3  1.4  1.5  1.6  1.7  1.8  1.9  2.0
 [92]  2.1  2.2  2.3  2.4  2.5  2.6  2.7  2.8  2.9  3.0  3.1  3.2  3.3
[105]  3.4  3.5  3.6  3.7  3.8  3.9  4.0  4.1  4.2  4.3  4.4  4.5  4.6
[118]  4.7  4.8  4.9  5.0  5.1  5.2  5.3  5.4  5.5  5.6  5.7  5.8  5.9
[131]  6.0  6.1  6.2  6.3  6.4  6.5  6.6  6.7  6.8  6.9  7.0
>
```

```
acc2       91.970802919708
conf_tab…  'table' int [1:2, 1:2] 2…
conf_tab…  'table' int [1:2, 1:2] 2…
idx        int [1:961] 2 3 5 7 8 9 …
m          961
n          1372L
var_wave   num [1:141] -7 -6.9 -6.8…
```

So, basically if I just do that now you can see that it just creating values between minus 7 to 7 with a difference of 0.1 ok, it just created these values.

(Refer Slide Time: 25:28)

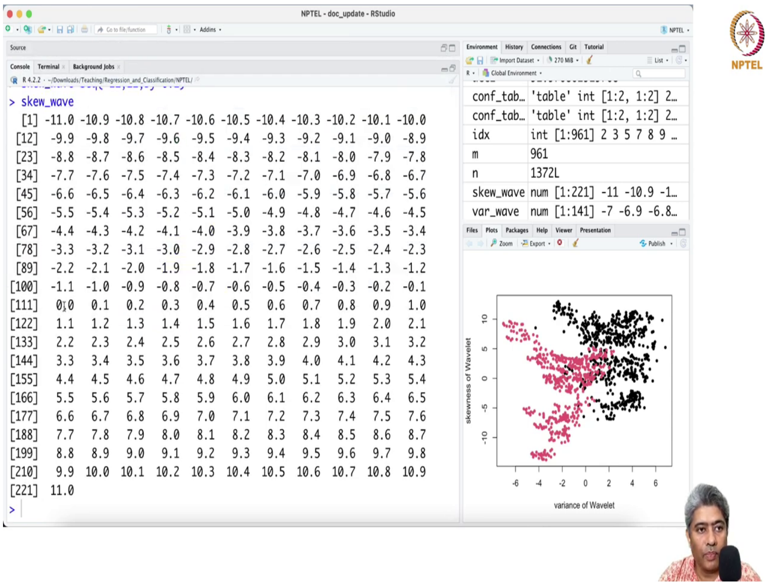


```r
acc2-acc1

## visualisation to understand the effect of feature engineering

test_data = dum_var = data.frame(matrix(NA,nrow=1,ncol=2))
colnames(test_data)=colnames(dum_var)=c("var_wave","skew_wave")


var_wave=seq(-7,7,by=0.1)
skew_wave=seq(-11,11,by=0.1)

```

```
conf_tab...  'table' int [1:2, 1:2] 2...
conf_tab...  'table' int [1:2, 1:2] 2...
idx          int [1:961] 2 3 5 7 8 9 ...
m            961
n            1372L
skew_wave    num [1:221] -11 -10.9 -1...
var_wave     num [1:141] -7 -6.9 -6.8...
```

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
> skew_wave=seq(-11,11,by=0.1)
>
```

(Refer Slide Time: 25:44)

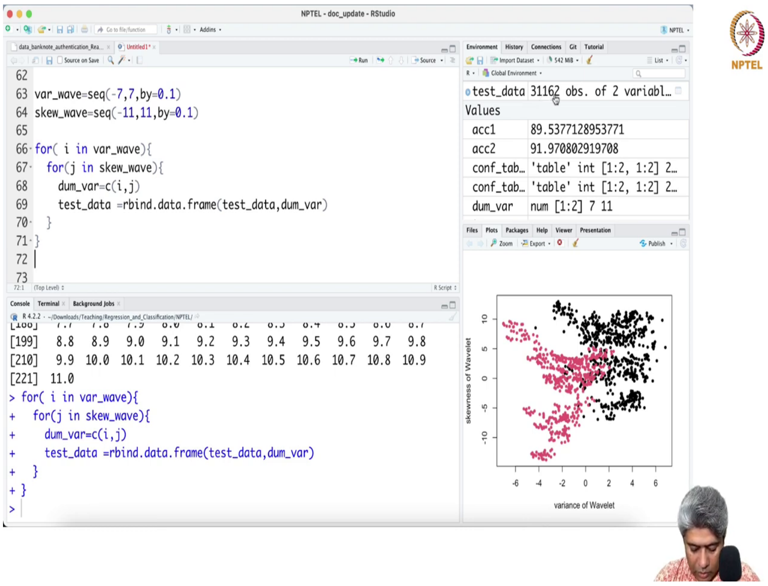


```
> skew_wave
  [1] -11.0 -10.9 -10.8 -10.7 -10.6 -10.5 -10.4 -10.3 -10.2 -10.1 -10.0
 [12]  -9.9  -9.8  -9.7  -9.6  -9.5  -9.4  -9.3  -9.2  -9.1  -9.0  -8.9
 [23]  -8.8  -8.7  -8.6  -8.5  -8.4  -8.3  -8.2  -8.1  -8.0  -7.9  -7.8
 [34]  -7.7  -7.6  -7.5  -7.4  -7.3  -7.2  -7.1  -7.0  -6.9  -6.8  -6.7
 [45]  -6.6  -6.5  -6.4  -6.3  -6.2  -6.1  -6.0  -5.9  -5.8  -5.7  -5.6
 [56]  -5.5  -5.4  -5.3  -5.2  -5.1  -5.0  -4.9  -4.8  -4.7  -4.6  -4.5
 [67]  -4.4  -4.3  -4.2  -4.1  -4.0  -3.9  -3.8  -3.7  -3.6  -3.5  -3.4
 [78]  -3.3  -3.2  -3.1  -3.0  -2.9  -2.8  -2.7  -2.6  -2.5  -2.4  -2.3
 [89]  -2.2  -2.1  -2.0  -1.9  -1.8  -1.7  -1.6  -1.5  -1.4  -1.3  -1.2
[100]  -1.1  -1.0  -0.9  -0.8  -0.7  -0.6  -0.5  -0.4  -0.3  -0.2  -0.1
[111]   0.0   0.1   0.2   0.3   0.4   0.5   0.6   0.7   0.8   0.9   1.0
[122]   1.1   1.2   1.3   1.4   1.5   1.6   1.7   1.8   1.9   2.0   2.1
[133]   2.2   2.3   2.4   2.5   2.6   2.7   2.8   2.9   3.0   3.1   3.2
[144]   3.3   3.4   3.5   3.6   3.7   3.8   3.9   4.0   4.1   4.2   4.3
[155]   4.4   4.5   4.6   4.7   4.8   4.9   5.0   5.1   5.2   5.3   5.4
[166]   5.5   5.6   5.7   5.8   5.9   6.0   6.1   6.2   6.3   6.4   6.5
[177]   6.6   6.7   6.8   6.9   7.0   7.1   7.2   7.3   7.4   7.5   7.6
[188]   7.7   7.8   7.9   8.0   8.1   8.2   8.3   8.4   8.5   8.6   8.7
[199]   8.8   8.9   9.0   9.1   9.2   9.3   9.4   9.5   9.6   9.7   9.8
[210]   9.9  10.0  10.1  10.2  10.3  10.4  10.5  10.6  10.7  10.8  10.9
[221]  11.0
>
```

conf_tab…	'table' int [1:2, 1:2] 2…
conf_tab…	'table' int [1:2, 1:2] 2…
idx	int [1:961] 2 3 5 7 8 9 …
m	961
n	1372L
skew_wave	num [1:221] -11 -10.9 -1…
var_wave	num [1:141] -7 -6.9 -6.8…

Similarly, if I just say skewness of wave is going to take sequence of value between minus 11 to 11 by 0.1 difference this is also going to create values between minus 11 and 11 with a difference of 0.1, ok equal distance of 0.1. So, I am getting grid I just created grid.

(Refer Slide Time: 26:01)



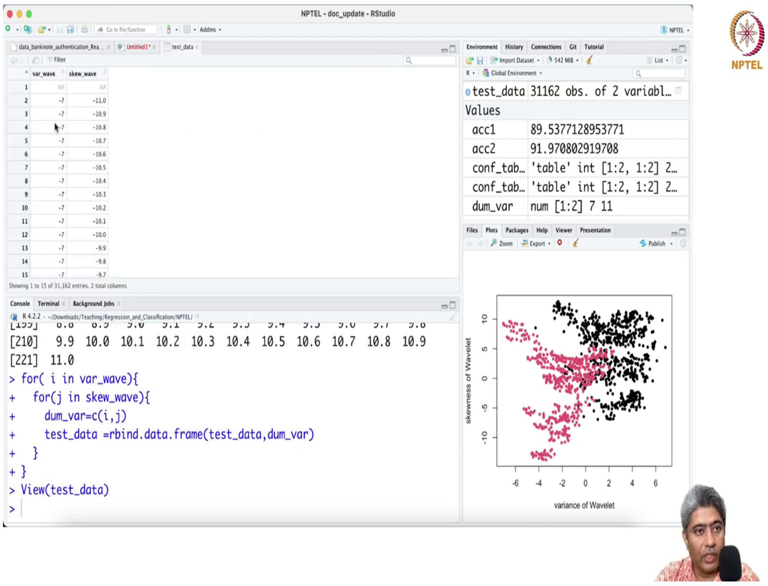
```r

var_wave=seq(-7,7,by=0.1)
skew_wave=seq(-11,11,by=0.1)

for( i in var_wave){
  for(j in skew_wave){
    dum_var=c(i,j)
    test_data =rbind.data.frame(test_data,dum_var)
  }
}


```

```
test_data   31162 obs. of 2 variabl…
Values
acc1        89.5377128953771
acc2        91.970802919708
conf_tab…   'table' int [1:2, 1:2] 2…
conf_tab…   'table' int [1:2, 1:2] 2…
dum_var     num [1:2] 7 11
```

```
[188]   7.7   7.8   7.9   8.0   8.1   8.2   8.3   8.4   8.5   8.6   8.7
[199]   8.8   8.9   9.0   9.1   9.2   9.3   9.4   9.5   9.6   9.7   9.8
[210]   9.9  10.0  10.1  10.2  10.3  10.4  10.5  10.6  10.7  10.8  10.9
[221]  11.0
> for( i in var_wave){
+   for(j in skew_wave){
+     dum_var=c(i,j)
+     test_data =rbind.data.frame(test_data,dum_var)
+   }
+ }
>
```

Now, I am going to create a joint grid; joint grid how that is how. So, for i in 1 i in variance of wave for j in skewness of wave and then dum variable equal to c of whatever i and j values are there in the thing and then what I am going to do? Test data equal to rbind data dot frame test data comma dummy variable, alright.

(Refer Slide Time: 27:13)



NPTEL - doc_update - RStudio

data_banknote_authentication_Rea... | Untitled1* | test_data

Showing 1 to 15 of 31,162 entries, 2 total columns

	var_wave	skew_wave
1	NA	NA
2	-7	-11.0
3	-7	-10.9
4	-7	-10.8
5	-7	-10.7
6	-7	-10.6
7	-7	-10.5
8	-7	-10.4
9	-7	-10.3
10	-7	-10.2
11	-7	-10.1
12	-7	-10.0
13	-7	-9.9
14	-7	-9.8
15	-7	-9.7

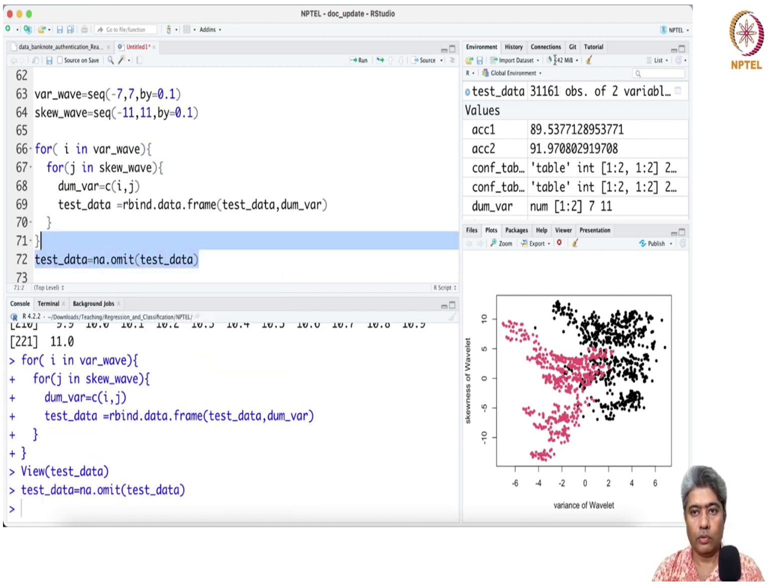
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/

```
[199]   8.8   8.9   9.0   9.1   9.2   9.3   9.4   9.5   9.6   9.7   9.8
[210]   9.9  10.0  10.1  10.2  10.3  10.4  10.5  10.6  10.7  10.8  10.9
[221]  11.0
> for( i in var_wave){
+   for(j in skew_wave){
+     dum_var=c(i,j)
+     test_data =rbind.data.frame(test_data,dum_var)
+   }
+ }
> View(test_data)
> 
```

test_data	31162 obs. of 2 variabl…
Values	
acc1	89.5377128953771
acc2	91.970802919708
conf_tab…	'table' int [1:2, 1:2] 2…
conf_tab…	'table' int [1:2, 1:2] 2…
dum_var	num [1:2] 7 11

So, let me just run it, ok. So, it has created test data with 31,162 observations if I just create that you will see that it has created all a complete grid of two values.

(Refer Slide Time: 27:27)



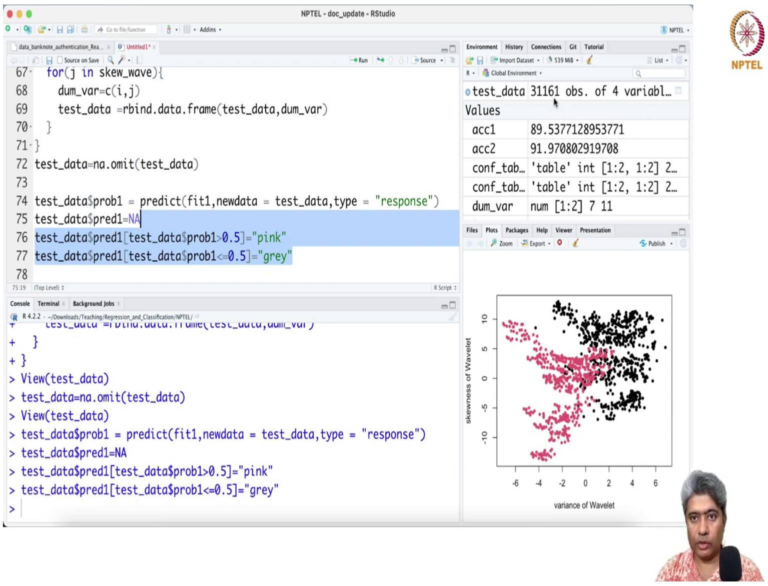
```r
var_wave=seq(-7,7,by=0.1)
skew_wave=seq(-11,11,by=0.1)

for( i in var_wave){
  for(j in skew_wave){
    dum_var=c(i,j)
    test_data =rbind.data.frame(test_data,dum_var)
  }
}
test_data=na.omit(test_data)
```

```
test_data  31161 obs. of 2 variabl…
Values
acc1       89.5377128953771
acc2       91.970802919708
conf_tab…  'table' int [1:2, 1:2] 2…
conf_tab…  'table' int [1:2, 1:2] 2…
dum_var    num [1:2] 7 11
```

```
[210]   9.9  10.0  10.1  10.2  10.3  10.4  10.5  10.6  10.7  10.8  10.9
[221]  11.0
> for( i in var_wave){
+   for(j in skew_wave){
+     dum_var=c(i,j)
+     test_data =rbind.data.frame(test_data,dum_var)
+   }
+ }
> View(test_data)
> test_data=na.omit(test_data)
>
```

And now what I am going to do I am going to create a let me just put a na dot omit dot omit. So, there will be the first row will be deleted and now you can see all the values are being created.

(Refer Slide Time: 27:48)



```
  for(j in skew_wave){
    dum_var=c(i,j)
    test_data =rbind.data.frame(test_data,dum_var)
  }
}
test_data=na.omit(test_data)

test_data$prob1 = predict(fit1,newdata = test_data,type = "response")
test_data$pred1=NA
test_data$pred1[test_data$prob1>0.5]="pink"
test_data$pred1[test_data$prob1<=0.5]="grey"

```

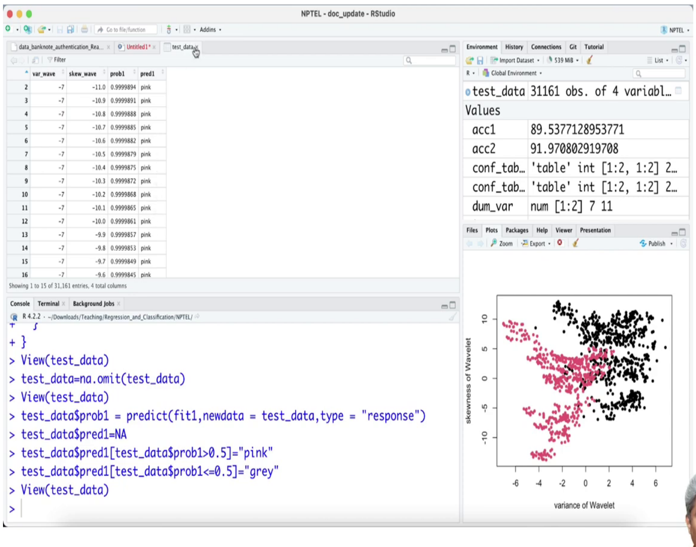
```
+     test_data =rbind.data.frame(test_data,dum_var)
+   }
+ }
> View(test_data)
> test_data=na.omit(test_data)
> View(test_data)
> test_data$prob1 = predict(fit1,newdata = test_data,type = "response")
> test_data$pred1=NA
> test_data$pred1[test_data$prob1>0.5]="pink"
> test_data$pred1[test_data$prob1<=0.5]="grey"
> 
```

```
test_data  31161 obs. of 4 variabl…
Values
  acc1       89.5377128953771
  acc2       91.970802919708
  conf_tab…  'table' int [1:2, 1:2] 2…
  conf_tab…  'table' int [1:2, 1:2] 2…
  dum_var    num [1:2] 7 11
```

Now, essentially what I am going to do is I am going to create for these all possible values I am going to create probability1 equal to predict fit from the 1st model newdata equal to test data and type equal to “response”, ok.

Now, what I am going to do is test data from the test data I am going to do a prediction and here is the whole thing the prediction is basically probability1 if it is greater than probability1 greater than 0.5 then the prediction will be “pink” basically it is a forged; that means, its forged, ok and if it is less than equal to 5 we will use grey; that means, it is not “grey” color I am just going to make this things, ok. Now, if you just go there you see pink and grey.

(Refer Slide Time: 29:37)



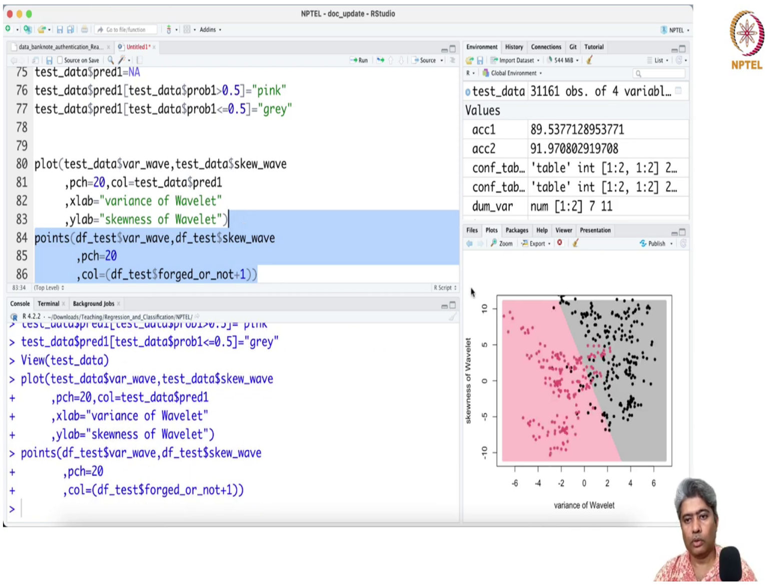
	var_wave	skew_wave	prob1	pred1
2	-7	-11.0	0.9999894	pink
3	-7	-10.9	0.9999891	pink
4	-7	-10.8	0.9999888	pink
5	-7	-10.7	0.9999885	pink
6	-7	-10.6	0.9999882	pink
7	-7	-10.5	0.9999879	pink
8	-7	-10.4	0.9999875	pink
9	-7	-10.3	0.9999872	pink
10	-7	-10.2	0.9999868	pink
11	-7	-10.1	0.9999865	pink
12	-7	-10.0	0.9999861	pink
13	-7	-9.9	0.9999857	pink
14	-7	-9.8	0.9999853	pink
15	-7	-9.7	0.9999849	pink
16	-7	-9.6	0.9999845	pink

Showing 1 to 15 of 31,161 entries, 4 total columns

```
R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/
+   }
+ }
> View(test_data)
> test_data=na.omit(test_data)
> View(test_data)
> test_data$prob1 = predict(fit1,newdata = test_data,type = "response")
> test_data$pred1=NA
> test_data$pred1[test_data$prob1>0.5]="pink"
> test_data$pred1[test_data$prob1<=0.5]="grey"
> View(test_data)
>
```

```
test_data  31161 obs. of 4 variabl…
Values
  acc1       89.5377128953771
  acc2       91.970802919708
  conf_tab…  'table' int [1:2, 1:2] 2…
  conf_tab…  'table' int [1:2, 1:2] 2…
  dum_var    num [1:2] 7 11
```

(Refer Slide Time: 29:44)



NPTEL - doc_update - RStudio

```
test_data$pred1=NA
test_data$pred1[test_data$prob1>0.5]="pink"
test_data$pred1[test_data$prob1<=0.5]="grey"


plot(test_data$var_wave,test_data$skew_wave
      ,pch=20,col=test_data$pred1
      ,xlab="variance of Wavelet"
      ,ylab="skewness of Wavelet")
points(df_test$var_wave,df_test$skew_wave
        ,pch=20
        ,col=(df_test$forged_or_not+1))
```

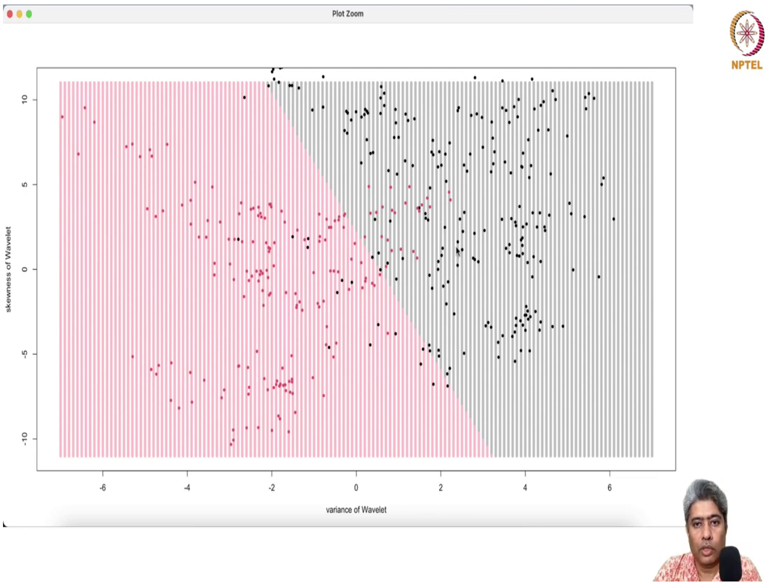
```
> test_data$pred1[test_data$prob1>0.5]="pink"
> test_data$pred1[test_data$prob1<=0.5]="grey"
> View(test_data)
> plot(test_data$var_wave,test_data$skew_wave
+       ,pch=20,col=test_data$pred1
+       ,xlab="variance of Wavelet"
+       ,ylab="skewness of Wavelet")
> points(df_test$var_wave,df_test$skew_wave
+         ,pch=20
+         ,col=(df_test$forged_or_not+1))
> 
```

```
test_data   31161 obs. of 4 variabl…
Values
acc1        89.5377128953771
acc2        91.970802919708
conf_tab…   'table' int [1:2, 1:2] 2…
conf_tab…   'table' int [1:2, 1:2] 2…
dum_var     num [1:2] 7 11
```

So, now what I am going to do? I am going to make a plot I am going to make a plot that test data dollar variance of wave test datadollarskewness of wave and then pch equal to 20 color equal to test data dollar my prediction color whatever the prediction color. Now from the above plot so, I can I will just take the lab labels xlabels and ylabels, ok.

So, that is the my predicted area. So, later the one which so, any point that will fall in this area will be called forged note and any points or any notes whose variance of wavelet and the skewness of the wavelet falls in this area will be called not forged or original note, ok. And then what I will do? I will I can just put the points essentially df testdollarvariance of wave comma df testdollarskewness of wave pch equal to 20 and color equal to df test dollar the forged or not plus 1.

(Refer Slide Time: 32:04)

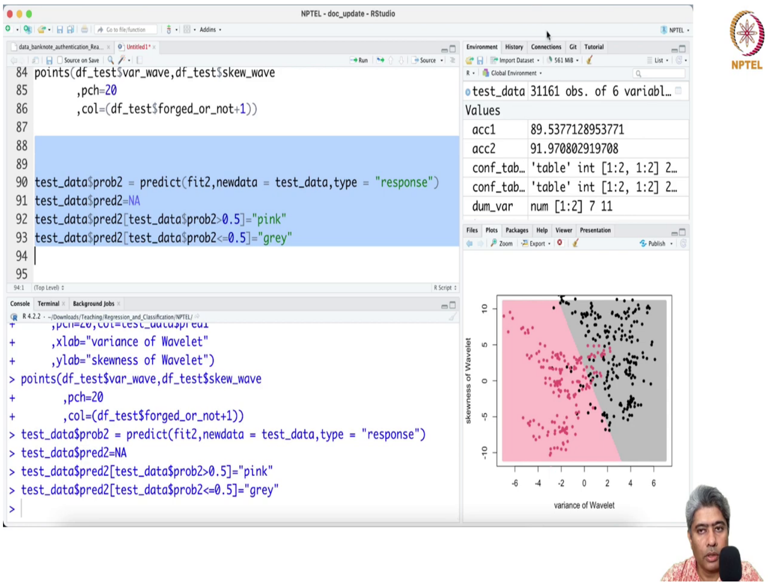


So, I have to just take this and then alright. So, if I just run this. So, now, if you see so, these are the points. So, let me just. So, clearly you can yeah I think this is now probably better right. Now, this looks good.

So, what we are seeing here that these are the points you can see there are some misclassification point the points which were definitely where will be misclassified and then there are like you know forged note, but they will be called as a original note and then there will be notes which are original, but they will be called as forged.

So, there is a effectively bit of a difficult this is sort of a difficult zone, ok. This is definitely a bit of a difficult zone, alright. Now, what I am going to do I am going to produce the same plot, but for the second model remember that second model were second model were developed using the engineered features with quadratic terms, alright.

(Refer Slide Time: 33:44)



```
points(df_test$var_wave,df_test$skew_wave
       ,pch=20
       ,col=(df_test$forged_or_not+1))



test_data$prob2 = predict(fit2,newdata = test_data,type = "response")
test_data$pred2=NA
test_data$pred2[test_data$prob2>0.5]="pink"
test_data$pred2[test_data$prob2<=0.5]="grey"

```

test_data 31161 obs. of 6 variabl...

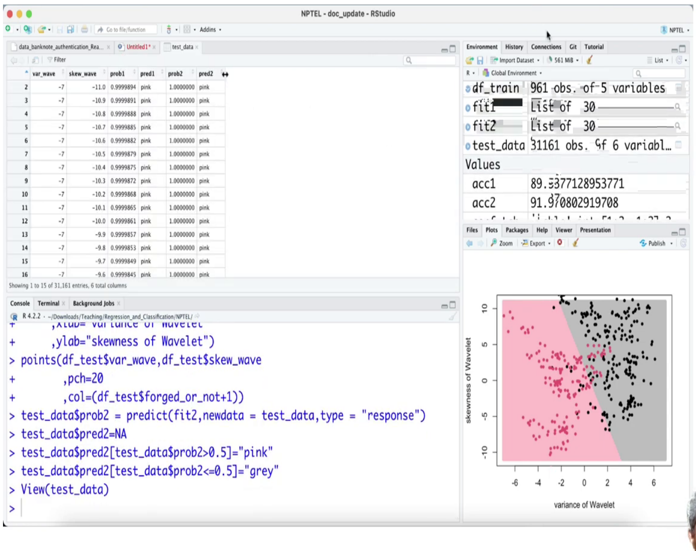
Values

acc1	89.5377128953771
acc2	91.970802919708
conf_tab...	'table' int [1:2, 1:2] 2...
conf_tab...	'table' int [1:2, 1:2] 2...
dum_var	num [1:2] 7 11

```
+       ,pch=20,col=test_data$pred1
+       ,xlab="variance of Wavelet"
+       ,ylab="skewness of Wavelet")
> points(df_test$var_wave,df_test$skew_wave
+        ,pch=20
+        ,col=(df_test$forged_or_not+1))
> test_data$prob2 = predict(fit2,newdata = test_data,type = "response")
> test_data$pred2=NA
> test_data$pred2[test_data$prob2>0.5]="pink"
> test_data$pred2[test_data$prob2<=0.5]="grey"
> 
```

So, I am going to copy this thing. So, probability2 with second model probability2 probability2 prediction2 probability2 prediction2 probability2 let me just run this.

(Refer Slide Time: 34:01)



NPTEL - doc_update - RStudio

data_banknote_authentication_Rea... | Untitled1* | test_data

	var_wave	skew_wave	prob1	pred1	prob2	pred2
2	-7	-11.0	0.9999894	pink	1.0000000	pink
3	-7	-10.9	0.9999891	pink	1.0000000	pink
4	-7	-10.8	0.9999888	pink	1.0000000	pink
5	-7	-10.7	0.9999885	pink	1.0000000	pink
6	-7	-10.6	0.9999882	pink	1.0000000	pink
7	-7	-10.5	0.9999879	pink	1.0000000	pink
8	-7	-10.4	0.9999875	pink	1.0000000	pink
9	-7	-10.3	0.9999872	pink	1.0000000	pink
10	-7	-10.2	0.9999868	pink	1.0000000	pink
11	-7	-10.1	0.9999865	pink	1.0000000	pink
12	-7	-10.0	0.9999861	pink	1.0000000	pink
13	-7	-9.9	0.9999857	pink	1.0000000	pink
14	-7	-9.8	0.9999853	pink	1.0000000	pink
15	-7	-9.7	0.9999849	pink	1.0000000	pink
16	-7	-9.6	0.9999845	pink	1.0000000	pink

Showing 1 to 15 of 31,161 entries, 6 total columns

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/

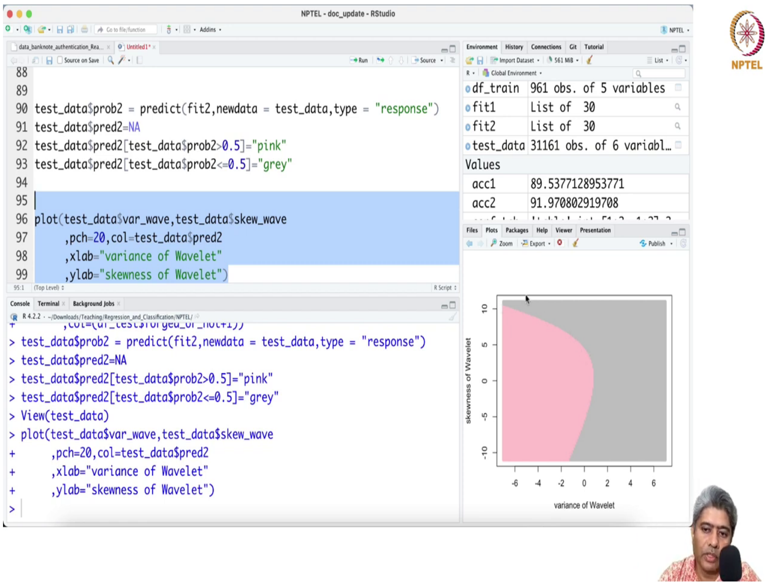
```
+       ,xlab="variance of Wavelet"
+       ,ylab="skewness of Wavelet")
> points(df_test$var_wave,df_test$skew_wave
+         ,pch=20
+         ,col=(df_test$forged_or_not+1))
> test_data$prob2 = predict(fit2,newdata = test_data,type = "response")
> test_data$pred2=NA
> test_data$pred2[test_data$prob2>0.5]="pink"
> test_data$pred2[test_data$prob2<=0.5]="grey"
> View(test_data)
> 
```

Global Environment

df_train	961 obs. of 5 variables
fit1	List of 30
fit2	List of 30
test_data	31161 obs. of 6 variabl…
Values	
acc1	89.3377128953771
acc2	91.970802919708

So, now if I have the test data, I have now got the second probability and their corresponding predictions as well.

(Refer Slide Time: 34:27)



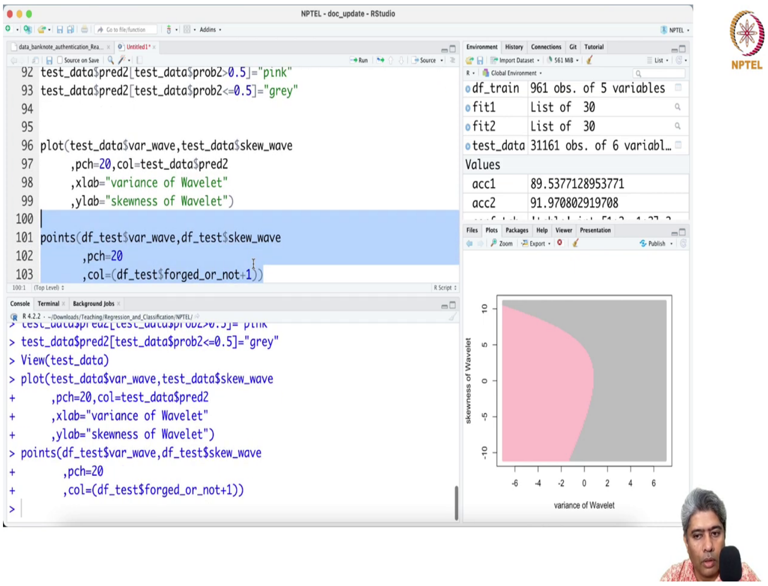
```
test_data$prob2 = predict(fit2,newdata = test_data,type = "response")
test_data$pred2=NA
test_data$pred2[test_data$prob2>0.5]="pink"
test_data$pred2[test_data$prob2<=0.5]="grey"


plot(test_data$var_wave,test_data$skew_wave
     ,pch=20,col=test_data$pred2
     ,xlab="variance of Wavelet"
     ,ylab="skewness of Wavelet")
```

```
> test_data$prob2 = predict(fit2,newdata = test_data,type = "response")
> test_data$pred2=NA
> test_data$pred2[test_data$prob2>0.5]="pink"
> test_data$pred2[test_data$prob2<=0.5]="grey"
> View(test_data)
> plot(test_data$var_wave,test_data$skew_wave
+       ,pch=20,col=test_data$pred2
+       ,xlab="variance of Wavelet"
+       ,ylab="skewness of Wavelet")
>
```

```
df_train    961 obs. of 5 variables
fit1        List of  30
fit2        List of  30
test_data   31161 obs. of 6 variabl…
Values
acc1        89.5377128953771
acc2        91.970802919708
```

So, now if I just run this, but the color will be instead of prediction1 the prediction2. Now if I just do that. So, now, you can see this is a completely different geometry a quadratic geometry sort of coming up nicely, ok sort of a ellipse or kind of thing behavior is coming up, which is a very interesting and now if we just put the points here, right.

(Refer Slide Time: 35:00)

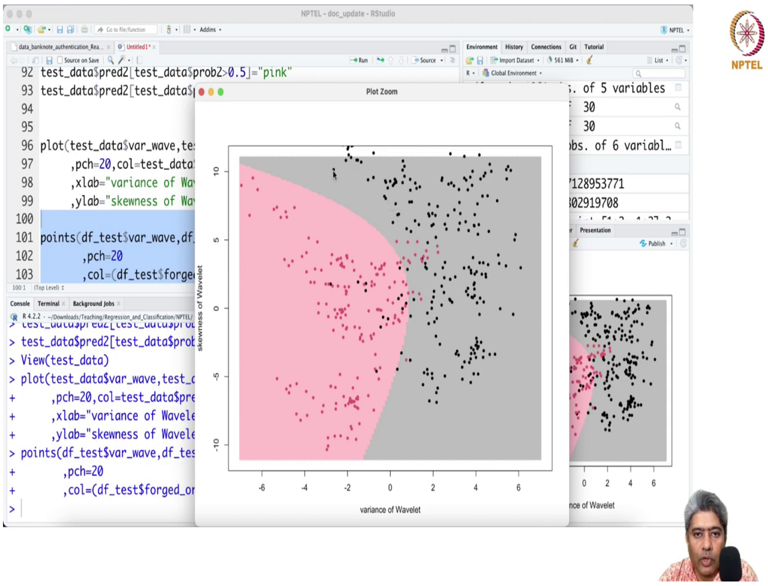


```r
test_data$pred2[test_data$prob2>0.5]="pink"
test_data$pred2[test_data$prob2<=0.5]="grey"


plot(test_data$var_wave,test_data$skew_wave
     ,pch=20,col=test_data$pred2
     ,xlab="variance of Wavelet"
     ,ylab="skewness of Wavelet")

points(df_test$var_wave,df_test$skew_wave
       ,pch=20
       ,col=(df_test$forged_or_not+1))
```

```
> test_data$pred2[test_data$prob2>0.5]="pink"
> test_data$pred2[test_data$prob2<=0.5]="grey"
> View(test_data)
> plot(test_data$var_wave,test_data$skew_wave
+      ,pch=20,col=test_data$pred2
+      ,xlab="variance of Wavelet"
+      ,ylab="skewness of Wavelet")
> points(df_test$var_wave,df_test$skew_wave
+        ,pch=20
+        ,col=(df_test$forged_or_not+1))
>
```

```
df_train   961 obs. of 5 variables
fit1       List of  30
fit2       List of  30
test_data  31161 obs. of 6 variabl…
Values
acc1       89.5377128953771
acc2       91.970802919708
```

(Refer Slide Time: 35:06)

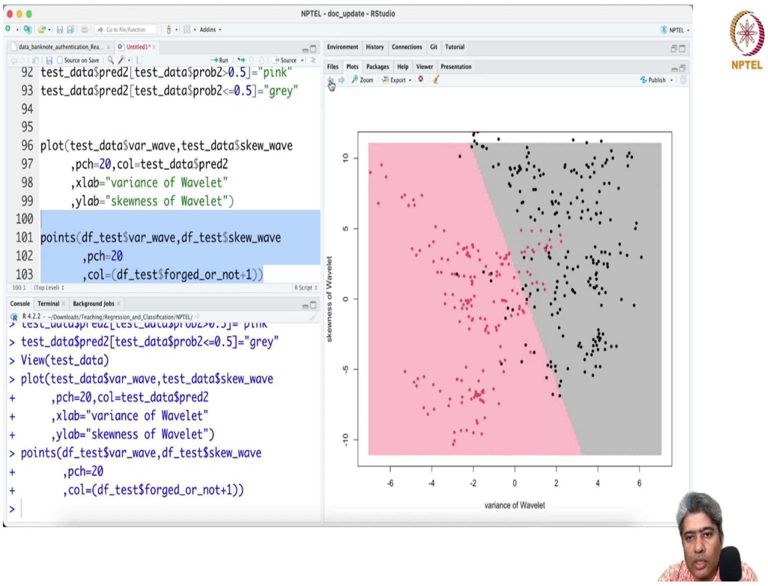


```
test_data$pred2[test_data$prob2>0.5]="pink"
test_data$pred2[test_data$


plot(test_data$var_wave,te
     ,pch=20,col=test_data
     ,xlab="variance of Wa
     ,ylab="skewness of Wa

points(df_test$var_wave,df
       ,pch=20
       ,col=(df_test$forge
```

```
> test_data$pred2[test_data$prob
> View(test_data)
> plot(test_data$var_wave,test_d
+      ,pch=20,col=test_data$pre
+      ,xlab="variance of Wavele
+      ,ylab="skewness of Wavele
> points(df_test$var_wave,df_tes
+        ,pch=20
+        ,col=(df_test$forged_or
>
```

So, put the points here then we can see that you know there is a it is trying to capture these points whereas. So, let me just you know little bit increase this guy. So, it will be let, yeah.

(Refer Slide Time: 35:30)



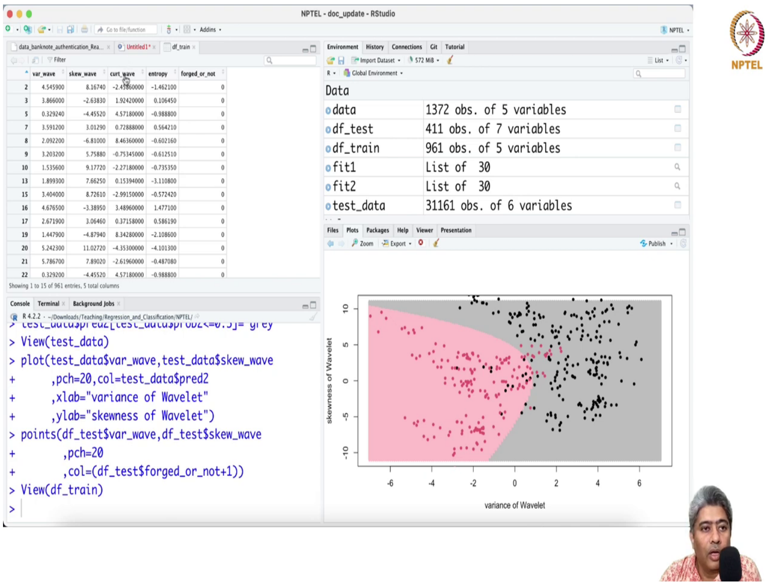
```
test_data$pred2[test_data$prob2>0.5]="pink"
test_data$pred2[test_data$prob2<=0.5]="grey"


plot(test_data$var_wave,test_data$skew_wave
      ,pch=20,col=test_data$pred2
      ,xlab="variance of Wavelet"
      ,ylab="skewness of Wavelet")

points(df_test$var_wave,df_test$skew_wave
        ,pch=20
        ,col=(df_test$forged_or_not+1))
```

```
> test_data$pred2[test_data$prob2>0.5]="pink"
> test_data$pred2[test_data$prob2<=0.5]="grey"
> View(test_data)
> plot(test_data$var_wave,test_data$skew_wave
+       ,pch=20,col=test_data$pred2
+       ,xlab="variance of Wavelet"
+       ,ylab="skewness of Wavelet")
> points(df_test$var_wave,df_test$skew_wave
+         ,pch=20
+         ,col=(df_test$forged_or_not+1))
> 
```

So, this is the quadratic behavior that we are seeing here and whereas, if you just this is the simple without any feature engineering, we will get a simple linear you know discriminator whereas, here we are getting a sort of a nice non-parametric quadratic discriminator. Here you can see these points are getting misclassified right, these points are getting misclassified whereas, here they are not getting misclassified this point is getting misclassified.

But it is better than too many and similarly I think if this point is getting misclassified here whereas, sorry about that, this is better this point are not getting misclassified. So, few points which are nicely not getting misclassified and trying to capture this and overall amount of the sample accuracy is going up.

(Refer Slide Time: 36:51)


NPTEL - doc_update - RStudio

data_banknote_authentication_Rea... | Untitled1* | df_train

	var_wave	skew_wave	curt_wave	entropy	forged_or_not
2	4.545900	8.16740	-2.45860000	-1.462100	0
3	3.866000	-2.63830	1.92420000	0.106450	0
5	0.329240	-4.45520	4.57180000	-0.988800	0
7	3.591200	3.01290	0.72888000	0.564210	0
8	2.092200	-6.81000	8.46360000	-0.602160	0
9	3.203200	5.75880	-0.75345000	-0.612510	0
10	1.535600	9.17720	-2.27180000	-0.735350	0
13	1.899300	7.66250	0.15394000	-3.110800	0
15	3.404000	8.72610	-2.99150000	-0.572420	0
16	4.676500	-3.38950	3.48960000	1.477100	0
17	2.671900	3.06460	0.37158000	0.586190	0
19	1.447900	-4.87940	8.34280000	-2.108600	0
20	5.242300	11.02720	-4.35300000	-4.101300	0
21	5.786700	7.89020	-2.61960000	-0.487080	0
22	0.329200	-4.45520	4.57180000	-0.988800	0

Showing 1 to 15 of 961 entries, 5 total columns

R 4.2.2 · ~/Downloads/Teaching/Regression_and_Classification/NPTEL/

```
> test_data$pred2[test_data$prob2<=0.5]="grey"
> View(test_data)
> plot(test_data$var_wave,test_data$skew_wave
+       ,pch=20,col=test_data$pred2
+       ,xlab="variance of Wavelet"
+       ,ylab="skewness of Wavelet")
> points(df_test$var_wave,df_test$skew_wave
+         ,pch=20
+         ,col=(df_test$forged_or_not+1))
> View(df_train)
>
```

Data

data	1372 obs. of 5 variables
df_test	411 obs. of 7 variables
df_train	961 obs. of 5 variables
fit1	List of 30
fit2	List of 30
test_data	31161 obs. of 6 variables

So, feature engineering typically helps remember that I have not tried the cubic and the other higher order transformations, in addition we have few more data for example, we have curtosis and entropy, which I do not have I have not used. So, my recommendation is you guys use that and check whether it helps you to improve the accuracy.

So, I will stop here and in the next video, I will try to do predictive analysis with a new data set with a new problem. I hope you are enjoying these hands-on real data analysis. Take care, bye.