

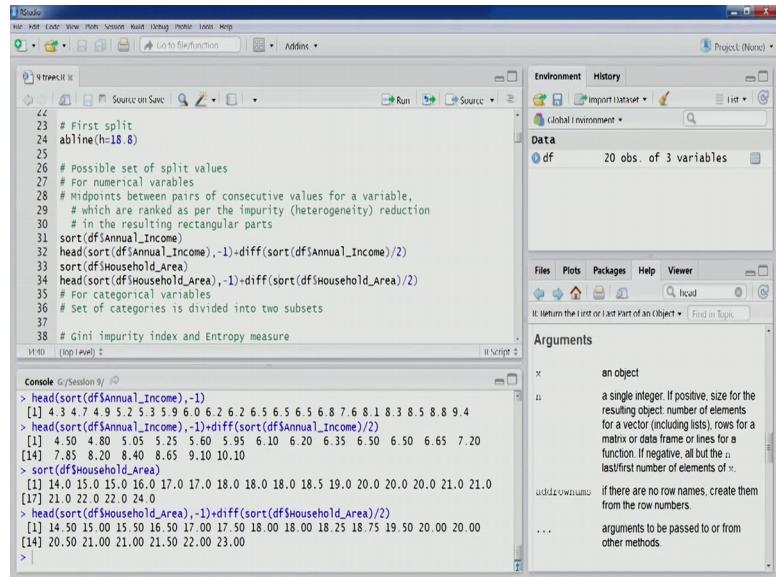
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

Lecture – 37
Classification and Regression Trees - Part II

Welcome to the course Business Analytics and Data Mining Modeling is Using R. So, in the previous lecture we started our discussion on classification and regression trees. So, we talked about two steps recursive partitioning and pruning and we started our discussion on recursive partitioning and we also started our exercise on the same and in the previous lecture. So, we were discussing about the possible set of split values that could be those values and how we can compute them, how we can get an idea about those split values using R.

So, in the previous lecture we talked about if the variable is numerical the predictor is numerical then what could be the possible set of split values. So, we talked about annual income and also household area that sedan car dataset that we are using for this exercise. So, we also computed midpoint values for these two variables and we talked about we have two variables and 19 midpoint values for each of them twenty observation we have in total. So, about 38 predictor value combination will have and out of these 38 predictor combination if the algorithm the implementation of that algorithm if it follows this process and out of these 38 combination we will have to select one optimal one which is going to reduce the impurity; that means, the heterogeneity.

(Refer Slide Time: 01:51)



The screenshot shows the RStudio interface. The code in the script editor is:

```
22 # First split
23 abline(h=18.8)
24
25 # Possible set of split values
26 # For numerical variables
27 # Midpoints between pairs of consecutive values for a variable,
28 # which are ranked as per the impurity (heterogeneity) reduction
29 # in the resulting rectangular parts
30 sort(df$Annual_Income)
31 head(sort(df$Annual_Income), -1)-diff(sort(df$Annual_Income)/2)
32 sort(df$Household_Area)
33 head(sort(df$Household_Area), -1)-diff(sort(df$Household_Area)/2)
34 # For categorical variables
35 # Set of categories is divided into two subsets
36
37 # Gini impurity index and Entropy measure
38
```

The console output shows the results of the sorting and midpoint calculations:

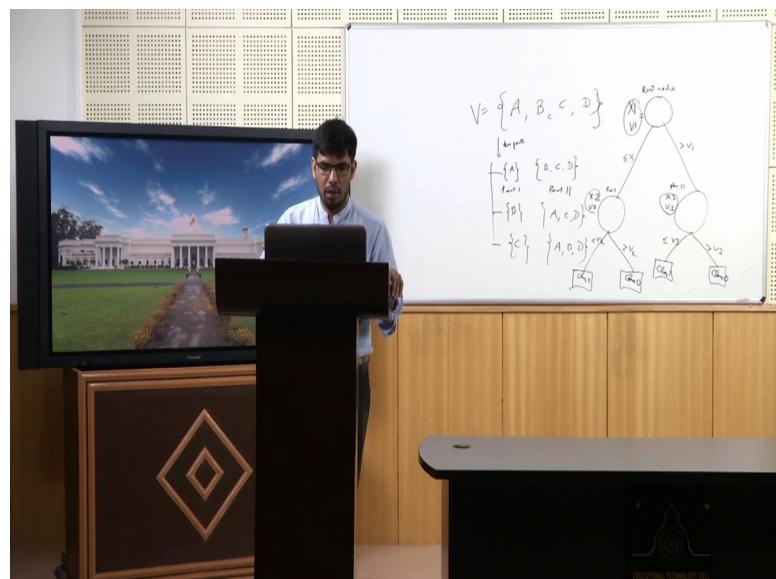
```
> head(sort(df$Annual_Income), -1)
[1] 4.3 4.7 4.9 5.2 5.3 5.9 6.0 6.2 6.5 6.5 6.8 7.6 8.1 8.3 8.5 8.8 9.4
> head(sort(df$Annual_Income), -1)-diff(sort(df$Annual_Income)/2)
[1] 4.50 4.80 5.05 5.25 5.60 5.95 6.10 6.20 6.35 6.50 6.50 6.65 7.20
[14] 7.85 8.20 8.40 8.65 9.10 10.10
> sort(df$Household_Area)
[1] 14.0 15.0 15.0 16.0 17.0 18.0 18.0 18.5 19.0 20.0 20.0 20.0 21.0 21.0
[17] 21.0 22.0 22.0 24.0
> head(sort(df$Household_Area), -1)-diff(sort(df$Household_Area)/2)
[1] 14.50 15.00 15.50 16.50 17.00 17.50 18.00 18.25 18.75 19.50 20.00 20.00
[14] 20.50 21.00 21.00 21.50 22.00 23.00
> |
```

The environment pane shows a data frame 'df' with 20 observations and 3 variables.

That could be there in the resulting partition. So, resulting partition having the least impurity; that means, more you know homogenous partition so that particular value combination would actually be selected for first spilt.

So, what if the variable if our variable is categorical? So, in that particular case the set of categories that we have they are divided into two subsets, for example, if we have a particular variable.

(Refer Slide Time: 02:29)



Let us say we have this variable. So, our values on the categories that are there, they could be this. So, from this we have to we can have many midpoint many set of possible candidates here right. So, there could be different you know value there different options here for example, you know this could be one. So, we have to create two parts from here. So, one category will go into one part the other categories will go into the other part right part 1 and part 2. So, in this fashion there could be various other candidates it could be B and others could be here then similarly it could be you know C and the others could be here. So, in this fashion there could be many combinations of these splits. So, there could be many split value the predictor and split value combination in this case also.

So, for categorical variable this is how we can create you know different combination of variable and split value. So, two subsets, for each the variable 4 categories A B C D, so all you know two subsets combination could be the different values that can be used as the possible set of candidate.

Now, let us talk about the impurity measures that we could be using for in this in this in this particular algorithm classification and regression tree. So, impurity measures that we are going to cover is two measures mainly a gini index and entropy measure. So, let say start our discussion on gini index.

(Refer Slide Time: 04:45)

CLASSIFICATION & REGRESSION TREES

- Impurity Measures
 - Gini index and Entropy measure
- Gini Index

For an outcome variable with m classes, Gini impurity index for a rectangular part is defined as

$$gini = 1 - \sum_{k=1}^m P_k^2$$

Where P_k is the proportion of rectangular part observations belonging to class k

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

6

So, for an, so both these measures whether gini index or entropy measures, they are in a sense major the impurity. So, for impurity for the original, original rectangle, original

group in our in our data and then once we create partitions. So, two parts part one and part two for each of those parts we can further compute the impurity using these matrix. So, then later on we can compare that the after we have done the particular partition after we have done a particular split whether there has been a decrease in impurity. So, how do we measure that impurity of different partitions? So, these are the two matrix which can be used gini index and entropy measure.

So, let us talk about the gini index first. So, for an outcome variable with m classes, gini impurity index for a rectangular part is defined as this $gini = 1 - \sum_{k=1}^m p_k^2$, where p_k is the proportion of rectangular part observation belonging to class k . So, for each class if we have m classes, for each class will have to compute the proportion of observations belonging to that class in that particular rectangular part.

So, for example, if we had the full original rectangle all the observations and. So, we can compute the you know for each class, class 1 to m c_1, c_2 up to c_m for each class we will have to compute the proportion values right proportion of observations belonging to class one in that particular rectangular part. Portion of observation belonging to class c_2 again in that same rectangular, in this fashion for all classes c_1 to c_m we will have to compute the this proportion values P_k and then square and summation of this. So, this will actually represent, this will actually the summation of this once we subtract this value from one this is actually going to represent the impurity right.

So, this will give us the impurity index for the rectangular part and once we create partition once we do a split we will have two more parts. So, for those two parts again we can use the same formula to compute their impurity value and these two parts we can add these two, we can add the impurity values of these two parts and then we can compare it with the original rectangular partition and see how much impurity has been reduced because of the partitioning alright. So, this is one particular metric that we can use.

Let us talk about the second metric entropy measure. So, before that let us understand the values gini values range. So, gini values lie in this range $0 \leq gini \leq 1$.

(Refer Slide Time: 08:04)

CLASSIFICATION & REGRESSION TREES

- Gini Index
 - Gini values lie in the range $\{0, (m-1)/m\}$ for m-class scenario and $\{0, 0.5\}$ for two-class scenario
- Entropy Measure
 - For an outcome variable with m classes, entropy for a rectangular part is defined as

$$\text{Entropy} = - \sum_{k=1}^m P_k \log_2(P_k)$$

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

7

So, if there are m classes. So, this is going to be the range for gini index and if there are if there are just two classes so the range is going to be 0 to 0.5. So, for 0 to 0 0.5, when the how we compute these two range? When in a two class scenario if the representation of both the classes is equal right, in that case the proportion would be 0.5 and 0.5 for both the classes. Now, if you go back to the expression here 1 minus summation over P_k square. So, we have use 0.5 and 0.5 for both the classes. So, you would get that value right. So, the value that you get and then that is going to be 0.5 so that is going to be the highest value. So, when we have the equal representation from all the classes the value the gini index value is going to be the highest because there that is the situation where the impurity is, where the impurity is highest because the observations belonging to different classes they are equal. If there in a particular rectangular partition if most of the observations belong to one particular class then of course, impurity is less because very few observation would be belonging to other class.

If this you know this particular ratio keep on decreasing and becomes equal where you know the different classes the observation belong different classes they are in equal proportion then of course, the impurity is going to be the highest and that is also you know indicated in this particular range. So, m class scenario the value is going to be $0 m$ minus one divided by m and 2 class scenario the value the range is going to be 0 to 0 0.5.

Let us talk about the next metric that is entropy measure. So, for an outcome variable with m classes and entropy for a rectangular part is defined as this entropy minus summation over k equal to 1 to m and $P_k \log_2 P_k$. So, this is how we compute the entropy value. So, as we discussed for gini index right same thing P_k stands for the same thing proportion of class k members in the rectangle in the rectangular part. So, then we compute that value then we take log of it log base 2 of it and then multiply these value and then sum it over all classes and the minus of that is going to be the entropy value.

(Refer Slide Time: 10:53)

CLASSIFICATION & REGRESSION TREES

- Entropy Measure
 - Entropy values lie in the range $\{0, \log_2(m)\}$ for m -class scenario and $\{0, 1\}$ for two-class scenario
- Open RStudio
- Tree diagram or tree structure
 - Each split of p -dimensional space into two parts can be depicted as a split of a node in a decision tree into two child nodes
 - First split creates branches of root node

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

8

So, the range for entropy value, here it is going to be 0 and $\log m$ base 2 for m class scenario and 0 and 1 for 2 class scenario, how?

So, for example, for two class scenario the highest impurity is going to be in this situation when the members belonging to each of those two classes they are in equal proportion they are in equal numbers. So, in that case the P_k value is going to be 1 by 2 or 0.5. So, if the P_k value is 1 by 2 you can plug in that value in this particular expression and you will get that $\log_2(1/2)$ is going to be you know minus 1. So, that minus they will cancel out and then P_k is there then that you will get the 1 by 2 and then for the second the other class also it will compute this value and once you sum it 1 by 2 plus 1 by 2 is going to be 1. So, this is how the range is.

So, highest impurity highest impurity scenario is when all the classes they have equal proportion they have equal representation in a particular rectangular part right. So, that is when the highest impurity is going to be there and that will also give us the range for entropy values and also for gini index. So, what we will do? To understand more about these two particular matrix will do a simple exercise in R. So, let us go back.

(Refer Slide Time: 12:38)

The screenshot shows the RStudio interface. The left pane displays an R script with the following code:

```

1 sort(df$Annual_Income)
2 head(sort(df$Annual_Income), -1) + diff(sort(df$Annual_Income)/2)
3 sort(df$Household_Area)
4 head(sort(df$Household_Area), -1) + diff(sort(df$Household_Area)/2)
5 # For categorical variables
6 # Set of categories is divided into two subsets
7
8 # Gini Impurity index and Entropy measure
9 # plot of gini vs. P1 (proportion of observations in class 1)
10 # for a two-class case
11 P1<-seq(0,1,0.1)
12 gini<-NULL
13 for(i in 1:length(P1)) {
14   gini[i]<-P1[i]^2 + (1-P1[i])^2
15 }
16 plot(P1,gini, ylab = "Gini index", type = "l")
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
669
670
671
672
673
674
675
676
677
678
679
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
787
788
789
789
790
791
792
793
794
795
796
797
797
798
799
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
817
818
819
819
820
821
822
823
824
825
826
827
827
828
829
829
830
831
832
833
834
835
836
837
837
838
839
839
840
841
842
843
844
845
846
847
847
848
849
849
850
851
852
853
854
855
856
857
857
858
859
859
860
861
862
863
864
865
866
866
867
868
868
869
869
870
871
872
873
874
875
876
876
877
878
878
879
879
880
881
882
883
884
885
886
886
887
888
888
889
889
890
891
892
893
894
895
895
896
896
897
897
898
898
899
899
900
901
902
903
904
905
905
906
907
907
908
909
909
910
911
912
913
913
914
914
915
915
916
916
917
917
918
918
919
919
920
920
921
921
922
922
923
923
924
924
925
925
926
926
927
927
928
928
929
929
930
930
931
931
932
932
933
933
934
934
935
935
936
936
937
937
938
938
939
939
940
940
941
941
942
942
943
943
944
944
945
945
946
946
947
947
948
948
949
949
950
950
951
951
952
952
953
953
954
954
955
955
956
956
957
957
958
958
959
959
960
960
961
961
962
962
963
963
964
964
965
965
966
966
967
967
968
968
969
969
970
970
971
971
972
972
973
973
974
974
975
975
976
976
977
977
978
978
979
979
980
980
981
981
982
982
983
983
984
984
985
985
986
986
987
987
988
988
989
989
990
990
991
991
992
992
993
993
994
994
995
995
996
996
997
997
998
998
999
999
1000
1000
1001
1001
1002
1002
1003
1003
1004
1004
1005
1005
1006
1006
1007
1007
1008
1008
1009
1009
1010
1010
1011
1011
1012
1012
1013
1013
1014
1014
1015
1015
1016
1016
1017
1017
1018
1018
1019
1019
1020
1020
1021
1021
1022
1022
1023
1023
1024
1024
1025
1025
1026
1026
1027
1027
1028
1028
1029
1029
1030
1030
1031
1031
1032
1032
1033
1033
1034
1034
1035
1035
1036
1036
1037
1037
1038
1038
1039
1039
1040
1040
1041
1041
1042
1042
1043
1043
1044
1044
1045
1045
1046
1046
1047
1047
1048
1048
1049
1049
1050
1050
1051
1051
1052
1052
1053
1053
1054
1054
1055
1055
1056
1056
1057
1057
1058
1058
1059
1059
1060
1060
1061
1061
1062
1062
1063
1063
1064
1064
1065
1065
1066
1066
1067
1067
1068
1068
1069
1069
1070
1070
1071
1071
1072
1072
1073
1073
1074
1074
1075
1075
1076
1076
1077
1077
1078
1078
1079
1079
1080
1080
1081
1081
1082
1082
1083
1083
1084
1084
1085
1085
1086
1086
1087
1087
1088
1088
1089
1089
1090
1090
1091
1091
1092
1092
1093
1093
1094
1094
1095
1095
1096
1096
1097
1097
1098
1098
1099
1099
1100
1100
1101
1101
1102
1102
1103
1103
1104
1104
1105
1105
1106
1106
1107
1107
1108
1108
1109
1109
1110
1110
1111
1111
1112
1112
1113
1113
1114
1114
1115
1115
1116
1116
1117
1117
1118
1118
1119
1119
1120
1120
1121
1121
1122
1122
1123
1123
1124
1124
1125
1125
1126
1126
1127
1127
1128
1128
1129
1129
1130
1130
1131
1131
1132
1132
1133
1133
1134
1134
1135
1135
1136
1136
1137
1137
1138
1138
1139
1139
1140
1140
1141
1141
1142
1142
1143
1143
1144
1144
1145
1145
1146
1146
1147
1147
1148
1148
1149
1149
1150
1150
1151
1151
1152
1152
1153
1153
1154
1154
1155
1155
1156
1156
1157
1157
1158
1158
1159
1159
1160
1160
1161
1161
1162
1162
1163
1163
1164
1164
1165
1165
1166
1166
1167
1167
1168
1168
1169
1169
1170
1170
1171
1171
1172
1172
1173
1173
1174
1174
1175
1175
1176
1176
1177
1177
1178
1178
1179
1179
1180
1180
1181
1181
1182
1182
1183
1183
1184
1184
1185
1185
1186
1186
1187
1187
1188
1188
1189
1189
1190
1190
1191
1191
1192
1192
1193
1193
1194
1194
1195
1195
1196
1196
1197
1197
1198
1198
1199
1199
1200
1200
1201
1201
1202
1202
1203
1203
1204
1204
1205
1205
1206
1206
1207
1207
1208
1208
1209
1209
1210
1210
1211
1211
1212
1212
1213
1213
1214
1214
1215
1215
1216
1216
1217
1217
1218
1218
1219
1219
1220
1220
1221
1221
1222
1222
1223
1223
1224
1224
1225
1225
1226
1226
1227
1227
1228
1228
1229
1229
1230
1230
1231
1231
1232
1232
1233
1233
1234
1234
1235
1235
1236
1236
1237
1237
1238
1238
1239
1239
1240
1240
1241
1241
1242
1242
1243
1243
1244
1244
1245
1245
1246
1246
1247
1247
1248
1248
1249
1249
1250
1250
1251
1251
1252
1252
1253
1253
1254
1254
1255
1255
1256
1256
1257
1257
1258
1258
1259
1259
1260
1260
1261
1261
1262
1262
1263
1263
1264
1264
1265
1265
1266
1266
1267
1267
1268
1268
1269
1269
1270
1270
1271
1271
1272
1272
1273
1273
1274
1274
1275
1275
1276
1276
1277
1277
1278
1278
1279
1279
1280
1280
1281
1281
1282
1282
1283
1283
1284
1284
1285
1285
1286
1286
1287
1287
1288
1288
1289
1289
1290
1290
1291
1291
1292
1292
1293
1293
1294
1294
1295
1295
1296
1296
1297
1297
1298
1298
1299
1299
1300
1300
1301
1301
1302
1302
1303
1303
1304
1304
1305
1305
1306
1306
1307
1307
1308
1308
1309
1309
1310
1310
1311
1311
1312
1312
1313
1313
1314
1314
1315
1315
1316
1316
1317
1317
1318
1318
1319
1319
1320
1320
1321
1321
1322
1322
1323
1323
1324
1324
1325
1325
1326
1326
1327
1327
1328
1328
1329
1329
1330
1330
1331
1331
1332
1332
1333
1333
1334
1334
1335
1335
1336
1336
1337
1337
1338
1338
1339
1339
1340
1340
1341
1341
1342
1342
1343
1343
1344
1344
1345
1345
1346
1346
1347
1347
1348
1348
1349
1349
1350
1350
1351
1351
1352
1352
1353
1353
1354
1354
1355
1355
1356
1356
1357
1357
1358
1358
1359
1359
1360
1360
1361
1361
1362
1362
1363
1363
1364
1364
1365
1365
1366
1366
1367
1367
1368
1368
1369
1369
1370
1370
1371
1371
1372
1372
1373
1373
1374
1374
1375
1375
1376
1376
1377
1377
1378
1378
1379
1379
1380
1380
1381
1381
1382
1382
1383
1383
1384
1384
1385
1385
1386
1386
1387
1387
1388
1388
1389
1389
1390
1390
1391
1391
1392
1392
1393
1393
1394
1394
1395
1395
1396
1396
1397
1397
1398
1398
1399
1399
1400
1400
1401
1401
1402
1402
1403
1403
1404
1404
1405
1405
1406
1406
1407
1407
1408
1408
1409
1409
1410
1410
1411
1411
1412
1412
1413
1413
1414
1414
1415
1415
1416
1416
1417
1417
1418
1418
1419
1419
1420
1420
1421
1421
1422
1422
1423
1423
1424
1424
1425
1425
1426
1426
1427
1427
1428
1428
1429
1429
1430
1430
1431
1431
1432
1432
1433
1433
1434
1434
1435
1435
1436
1436
1437
1437
1438
1438
1439
1439
1440
1440
1441
1441
1442
1442
1443
1443
1444
1444
1445
1445
1446
1446
1447
1447
1448
1448
1449
1449
1450
1450
1451
1451
1452
1452
1453
1453
1454
1454
1455
1455
1456
1456
1457
1457
1458
1458
1459
1459
1460
1460
1461
1461
1462
1462
1463
1463
1464
1464
1465
1465
1466
1466
1467
1467
1468
1468
1469
1469
1470
1470
1471
1471
1472
1472
1473
1473
1474
1474
1475
1475
1476
1476
1477
1477
1478
1478
1479
1479
1480
1480
1481
1481
1482
1482
1483
1483
1484
1484
1485
1485
1486
1486
1487
1487
1488
1488
1489
1489
1490
1490
1491
1491
1492
1492
1493
1493
1494
1494
1495
1495
1496
1496
1497
1497
1498
1498
1499
1499
1500
1500
1501
1501
1502
1502
1503
1503
1504
1504
1505
1505
1506
1506
1507
1507
1508
1508
1509
1509
1510
1510
1511
1511
1512
1512
1513
1513
1514
1514
1515
1515
1516
1516
1517
1517
1518
1518
1519
1519
1520
1520
1521
1521
1522
1522
1523
1523
1524
1524
1525
1525
1526
1526
1527
1527
1528
1528
1529
1529
1530
1530
1531
1531
1532
1532
1533
1533
1534
1534
1535
1535
1536
1536
1537
1537
1538
1538
1539
1539
1540
1540
1541
1541
1542
1542
1543
1543
1544
1544
1545
1545
1546
1546
1547
1547
1548
1548
1549
1549
1550
1550
1551
1551
1552
1552
1553
1553
1554
1554
1555
1555
1556
1556
1557
1557
1558
1558
1559
1559
1560
1560
1561
1561
1562
1562
1563
1563
1564
1564
1565
1565
1566
1566
1567
1567
1568
1568
1569
1569
1570
1570
1571
1571
1572
1572
1573
1573
1574
1574
1575
1575
1576
1576
1577
1577
1578
1578
1579
1579
1580
1580
1581
1581
1582
1582
1583
1583
1584
1584
1585
1585
1586
1586
1587
1587
1588
1588
1589
1589
1590
1590
1591
1591
1592
1592
1593
1593
1594
1594
1595
1595
1596
1596
1597
1597
1598
1598
1599
1599
1600
1600
1601
1601
1602
1602
1603
1603
1604
1604
1605
1605
1606
1606
1607
1607
1608
1608
1609
1609
1610
1610
1611
1611
1612
1612
1613
1613
1614
1614
1615
1615
1616
1616
1617
1617
1618
1618
1619
1619
1620
1620
1621
1621
1622
1622
1623
1623
1624
1624
1625
1625
1626
1626
1627
1627
1628
1628
1629
1629
1630
1630
1631
1631
1632
1632
1633
1633
1634
1
```

(Refer Slide Time: 13:20)

The screenshot shows the RStudio interface. The code in the script editor is:

```
51 sort(df$Annual_Income)
52 head(sort(df$Annual_Income),-1)-diff(sort(df$Annual_Income)/2)
53 sort(df$Household_Area)
54 head(sort(df$Household_Area),-1)-diff(sort(df$Household_Area)/2)
55 # For categorical variables
56 # Set of categories is divided into two subsets
57
58 # Gini impurity index and Entropy measure
59 # plot of gini vs. P1 (proportion of observations in class 1)
60 # for a two-class case
61 P1<-seq(0,1,0.1)
62 gini=NULL
63 for(i in 1:length(P1)) {
64   gini[i]=1-(P1[i]^2 + (1-P1[i])^2)
65 }
66 plot(P1,gini, ylab = "Gini index", type = "l")
67
```

The console output shows the sorted income and area data, followed by the calculated gini values:

```
[1] 4.50 4.80 5.05 5.25 5.60 5.95 6.10 6.20 6.35 6.50 6.65 7.20
[14] 7.85 8.20 8.40 8.65 9.10 10.10
> sort(df$Household_Area)
[1] 14.0 15.0 15.0 16.0 17.0 17.0 18.0 18.0 18.5 19.0 20.0 20.0 20.0 21.0 21.0
[14] 20.50 21.00 21.50 22.00 23.00
> head(sort(df$Household_Area),-1)-diff(sort(df$Household_Area)/2)
[1] 14.50 15.00 15.50 16.50 17.00 17.50 18.00 18.00 18.25 18.75 19.50 20.00 20.00
> P1<-seq(0,1,0.1)
> P1
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
>
```

The environment pane shows the global environment with the variable P1 defined.

So, against these proportion values P 1 values we are going to compute gini index values and then we are going to plot them. So, as we are already familiar with gini index formula. So, let us first initialize this gini variable. So, let us do the initialization and then we are going to run this loop i in 1 to length of P 1 that is eleven values in total. So, for each of those values for each of those proportion values we are going to compute the gini index. So, this was the, this is how we can express the gini index formula here in R, 1 minus and within parenthesis we have for each proportion we first, we use the proportion values and they take a square of it and then we do a sum and then we add all these values for the all the classes.

(Refer Slide Time: 14:32)

```

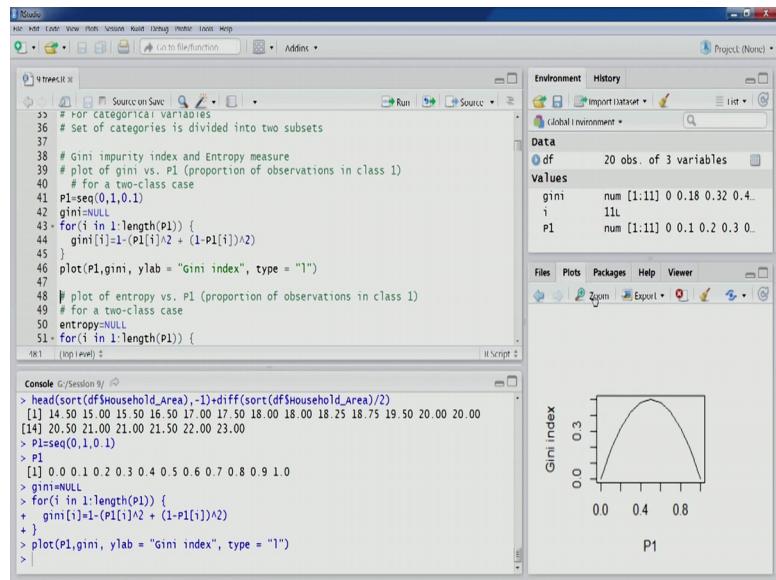
## Sort(df$Household_Area)
head(sort(df$Household_Area), -1) - diff(sort(df$Household_Area)) / 2
# For categorical variables
# Set of categories is divided into two subsets
# Gini impurity index and Entropy measure
# plot of gini vs. P1 (proportion of observations in class 1)
# for a two-class case
P1<-seq(0,1,0.1)
gini<-NULL
for(i in 1:length(P1)) {
  gini[i]=1-(P1[i])^2 + (1-P1[i])^2
}
plot(P1,gini, ylab = "Gini index", type = "l")
# plot of entropy vs. P1 (proportion of observations in class 1)
# for a two-class case

```

The screenshot shows the RStudio interface with the code in the script editor. The environment pane shows a data frame `df` with 20 observations and 3 variables. The `gini` variable is defined as a numeric vector of length 11. The arguments pane shows the definition of `gini` as an object.

So, let us compute this. We would see that a gini vector numeric vector has been created again 11 value. So, 11 gini index values corresponding to different proportion values right. So, let us plot this.

(Refer Slide Time: 14:47)



And this is the plot gini index versus proportion. Now, from here you can clearly understand as the proportion increases from 0 to 0.5 somewhere here you would see that gini index, index value is highest and it is 0.5 as we talked about right and as we further increase the increase the proportion right then again because P 1 this proportion

keeps on increasing again then the gini index value this will start decreasing right. So, this will keep on decreasing and again when the proportion is 1 this will go to this will become 0. So, this is how the values are going to be for gini metric, gini index.

The same thing we can do for entropy measure as well. So, let us plot and let us plot a graph entropy versus P_1 that is proportion of observation class 1. So, P_1 for we have already defined. So, let us initialize the entropy here now within for loop you can see how we have written the code for calculation of entropy value. So, you can see for each class we have one expression and each expression we have proportional and then multiplied by log base 2 value of that proportion and once we sum all these expression and we take a minus of it. So, let us run this loop to find out the entropy values you can see 11 values have been created right. You would see that first particular value that is n_a it is showing as $n_a n$, this is mainly because the proportion value for 0 here and log of 0 is not defined. So, because of that we have got this particular value.

So, let us plot. So, here you would see that in the plot function we are also using a spline function which will smooth smoothen the plot that we generate. So, let us see how what is going to happen. So, this is the plot here. So, you can see this particular plot is much more smoother than the plot that we had created for gini index. So, again here also as we move from 0 to 0.5 you would see that entropy measure is this particular value is maximum the value is 1 at 0.5 and as this proportion P_1 increases further this value goes down up to 0. So, this was about the two matrix, two matrix the gini index and entropy measure. So, let us talk further about our technique classification and regression trees.

So, next important point is the tree diagram or tree structure that we create. So, as we talked about the recursive partitioning steps. So, let us understand the tree diagram what is how this is going to be built. So, for each split of P dimensional space into two parts, so that is of course, the part of recursive partitioning, can be depicted as a split of a node in a decision tree into two child nodes. So, we can have a root node right, we can have a root node, let us this is our root node and this is the original party partition then the each split that we perform it can be denoted using two nodes here, right. So, this is one part 1, this is part 2. So, in this fashion they split that we are talking about can be created.

So, P dimensional space if it is P dimensional space we will start with the root node and this is going to be partition two parts are going to be created. So, this can be represented

in this fashion decision node having two child nodes. Now, once we have these two parts these two child nodes then again the same process would be applied on these two parts till you know, so the tree will start growing till the point we have created homogeneous partitions or homogeneous groups. So, first split creates branches of root node. So, as we can see. Now, two types of nodes in tree structure first one is a decision node. So, that is depicted with a circle here and then the second one is terminal or leaf node that is typically depicted using rectangle right.

So, these terminal nodes they typically they correspond to final rectangle parts. So, when we talk about just the recursive partitioning step where we build the full grown tree; that means, we get pure homogeneous parts. So, in that case we are going to have you know terminal nodes. So, for example, if this was you know root node and we created two partitions and once we created two partition we were able to achieve the homogeneous rectangles right. So, let us say further partitioning of this leads to homogeneous rectangles right. So, we will have, we can represent those nodes because they are going to be the terminal nodes leaf nodes using these rectangles right. So, these are decision nodes right. So, predictor and predictor value combination are going to be applied on these decision nodes and then the terminal nodes would indicate the actual class because this is now pure homogeneous group. So, it is going to be either class 1 or class 0, class 1 or class 0. So, in this fashion the tree structure could be there.

So, two types of node decision nodes and terminal node. So, decision nodes are the one where we apply the predictor value combinations and create a split and the terminal nodes or leaf nodes are the one where we finally, end up with pure homogeneous part homogeneous group and therefore, we can label it with the class name class 1 or class 0 if it is a two class case.

Now, let us understand the steps to classify new observations, new observation using tree based models. So, for a new observation once the tree has been built. So, new observation to be classified can be dropped down the tree. So, it can be dropped down from root node and then depending on the different comparison it will take different branches and the finally, it will end up with the terminal node or leaf node.

So, first step new observation to be classified is drop down the tree is starting from root node and at each decision node which also root, root node, root node is the first decision

node. So, at each decision node the appropriate branch is taken until we reach a leaf node right. So, for example, this is a variable, variable V 1 and you know let us say X 1, this is X 1 and then the corresponding value for this particular you know variable is V 1 and the split is created right. So, values less than V 1 they go this side values greater than V 1 they go this side two parts alright. So, in this fashion here again we will have another variable X 2 and the value V 2 here we will have X 3 and value V 3 right and then the observation having value less than V 2 will go here greater than V 2 will come here similarly for here.

So, in this fashion we will continue till we till the new observation reach the terminal node or leaf node where then finally, it is going to be classified as per the class of that particular terminal node. So, finally, at leaf node majority class is assigned to the new observation. So, now, this is going to be when we do not have any special class of interest where we are trying to maximize the overall accuracy or trying to minimize the overall misclassification error, but when we have a special class of interest as we have been talking about in previous lectures for other techniques the steps are going to change a bit. For example for a class of interest scenario proportion of records belonging to the class of interest is compared with the user specified cut off value for the same right.

(Refer Slide Time: 23:52)

CLASSIFICATION & REGRESSION TREES

- Steps to classify a new observation using tree based models
 - At leaf node, majority class is assigned to the new observation
 - For class of interest scenario, proportion of records belonging to the class of interest is compared with the user specified cut off value for the same
- Open RStudio

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

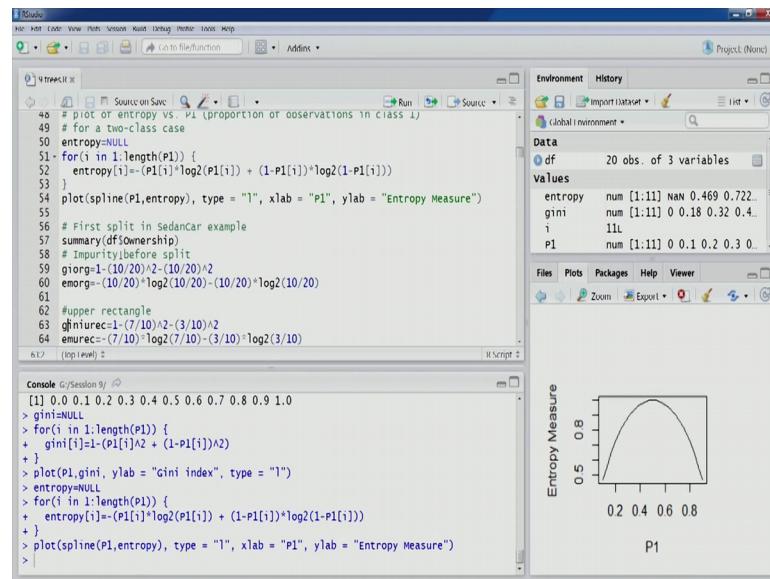
10

So, for the once you once we reach the leaf node typically you know when we talk about the recursive partitioning it is going to be a purely homogeneous partition. So, there is

going to be no such problem, but if the tree is not fully grown tree it has been pruned back pruning will discuss in coming lectures. So, in that case the partition the leaf terminal node might not be homogeneous and there could be some observation belonging to other classes. So, therefore, how do we decide? So, for when we try to, when we do not have any special class of interest and when we are looking to maximize overall accuracy in those situation we can just look at the majority class in the terminal node and assign that class to the new observation.

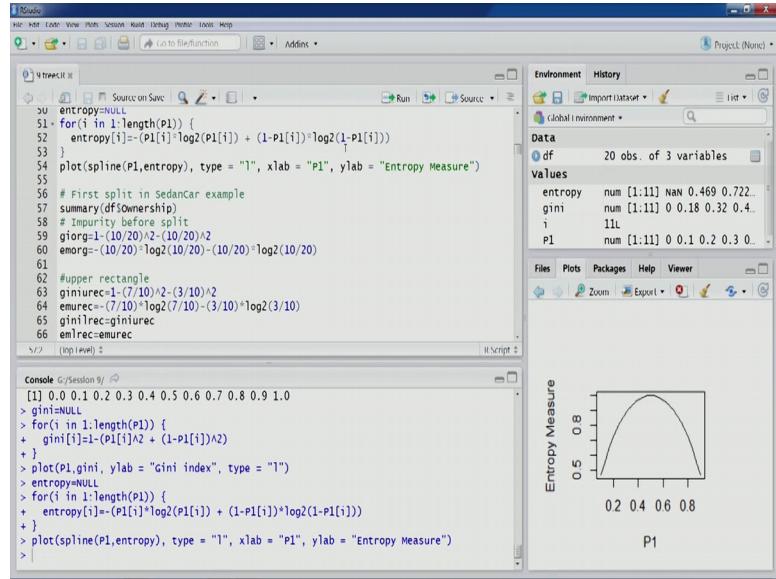
But when we have a class of interest we will compute the proportion of records belonging to that class of interest and then compare this particular proportion value to the user specified cut off value because that is the class of interest. So, we would like to identify more observations belonging to that class one even if it comes at the expense of miss identifying more observation belonging to other classes. So, the step is, this step final step is going to change depending on whether we have a class of interest or not.

(Refer Slide Time: 25:45)



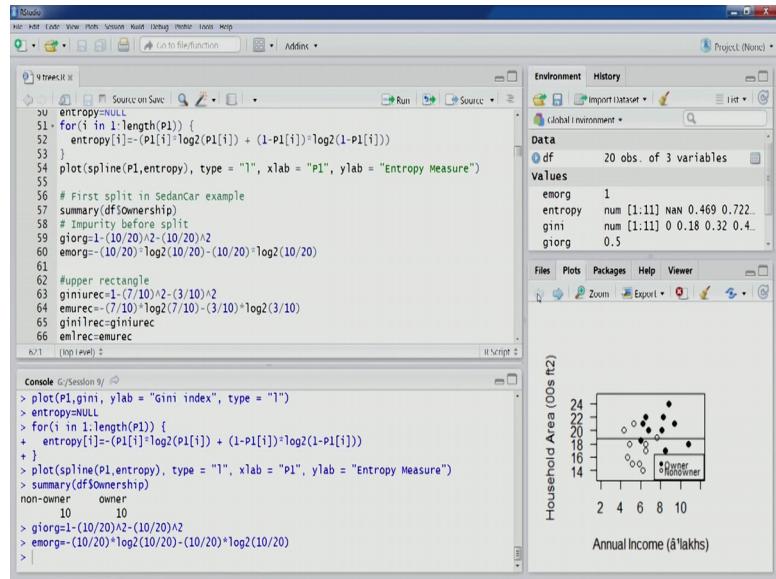
So, what we will do? I will go through a simple exercise in R. So, let us go back to R, but before that let us also go through and this exercise where we compute the impurity using two matrix that we talked about.

(Refer Slide Time: 26:03)



So, sedan car example that we have discussed before, let us look at the summary of this particular ownership variable. So, we have 10 observation belonging to non owner category and then observation belong to owner category. Now, the different matrix that we talked about the impurity index how we can compute. So, for gini index and entropy value for the original partition, original rectangle we can compute in this fashion you can see 1 minus because 10 observation belong to the non owner category out of 20. So, in this fashion we can compute the gini index for other classes as well. So, this would be the gini value. So, entropy value also we can compute in this fashion you can see 10 observation belong to owner non remaining 10 of belong to non owner. So, in this fashion we can compute the entropy value.

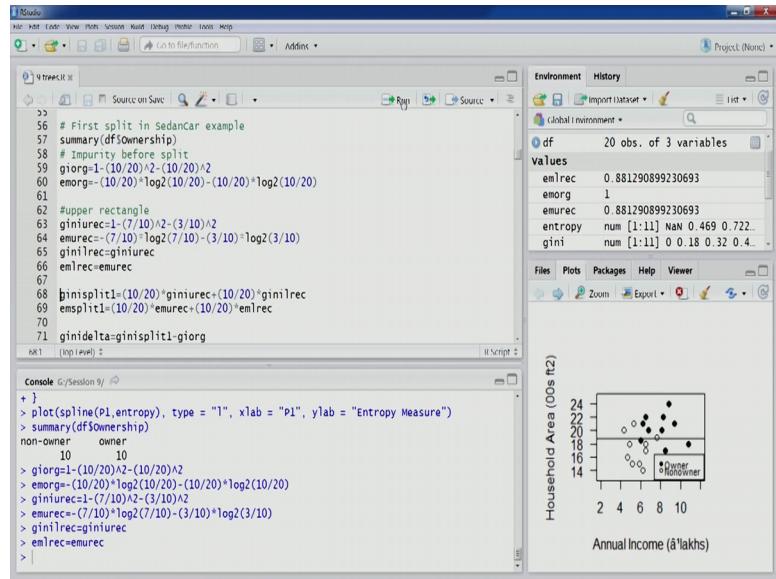
(Refer Slide Time: 26:50)



Now, once the first split that we had created earlier let us look at the graph. So, this was the graph you can see here we had created the first split at you know household area value of 18.8 and from this using this let us compute the gini entropy and entropy measure values. So, from this let us zoom into this particular plot. So, in the upper rectangular part you can see we have 7 observations belonging to the owner class and 3 observations belonging to the non owner class. So, it is 7 out of 10 to owner and 3 out of 10 non owner for upper rectangular part. So, gini for upper rectangular is going to be $1 - \frac{7}{10} \times \frac{3}{7}$ divided by 10 and that is square of that then $\frac{3}{10} \times \frac{7}{10}$ divided by 10 square of that. So, in this fashion we can compute the gini value for upper rectangular. Similarly for the entropy value for the upper rectangular also we can compute using similar approach. So, let us compute these two values.

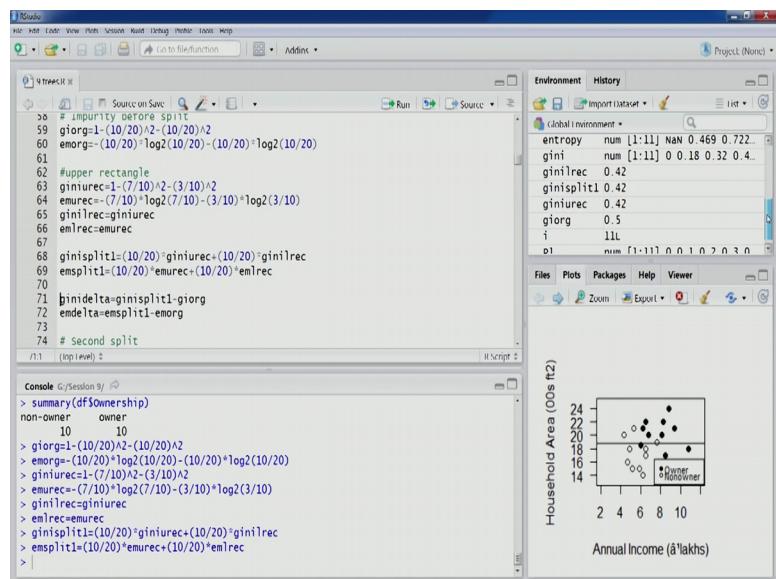
Now, if we look at the graph again you can see that lower rectangular part this is symmetric to the upper rectangular part in terms of proportion. So, portion of observations belonging to the owner and non owner. So, you know upper rectangular is dominated by owner lower rectangular is dominated by non owner, but the proportion they are very symmetric. So, the values for gini index and entropy measure they are going to be same. So, why not assign the same values for lower rectangular as well. So, gini value is going to be same as follow a rectangular is going to be same as upper rectangular. Similarly entropy value is going to be a follow rectangular, is going to be same as that for upper rectangular.

(Refer Slide Time: 28:51)



Once this is done, so for a split 1 we can compute the gini index value. So, we will add these two values for upper rectangular and lower rectangular. So, you can see we are also multiplying these value by their proportion here. So, 10 out of 20 observations in the upper rectangular, 10 out of 20 observation in the lower rectangular, this will give us the impurity index after first split and for entropy values of first split.

(Refer Slide Time: 29:27)

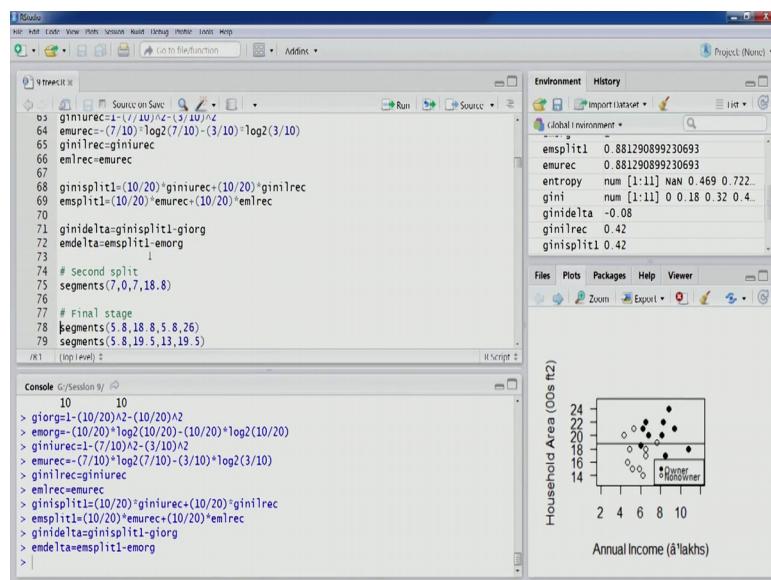


So, you can see in the environment section. So, values have been created you can split one around 0.88 and again you split on around 0.42, and the original values also you can

see original value is 0.5 g i o r g and e m o r g 1. So, now, we can compute the difference between you know that the delta that deduction that has happened in impurity. So, that is gini delta we can compute and e m delta. So, you can see e m delta minus this one minus 0.11 around minus 0.12 and gini delta is minus 0.08. So, if we can see there is a reduction in impurity. So, therefore, these two the first split is off course help us in achieving more, help us in achieving more homogeneous parts which is also very clearly visible from the plots as well.

So, in this fashion we can keep on continuing creating partition.

(Refer Slide Time: 30:38)



So, I will stop here and the other partition and the values gini values and the other exercises and discussion will continue in the next lecture.

Thank you.