

Machine learning for data analysis

Week 2 assignment

Running a random forest

I chose **addhealth** as my data set. General health (**H1GH1**) was used as response variable. It is a categorical variable with 5 levels (missing valued excluded). Therefore I collapsed it into a categorical variable with only two levels, and named this new variable as *health*. *Health* equal 1 means the participant is healthy, while *health* equal 0 means the participant is not healthy.

Then I run the random forest by **PROC HPFOREST** in SAS. Here is the output result.

Model Information		
Parameter	Value	
Variables to Try	3	(Default)
Maximum Trees	100	(Default)
Inbag Fraction	0.6	(Default)
Prune Fraction	0	(Default)
Prune Threshold	0.1	(Default)
Leaf Fraction	0.00001	(Default)
Leaf Size Setting	1	(Default)
Leaf Size Used	1	
Category Bins	30	(Default)
Interval Bins	100	
Minimum Category Size	5	(Default)
Node Size	100000	(Default)
Maximum Depth	20	(Default)
Alpha	1	(Default)
Exhaustive	5000	(Default)
Rows of Sequence to Skip	5	(Default)
Split Criterion	.	Gini
Preselection Method	.	Loh
Missing Value Handling	.	Valid value

A random selection of 3 explanatory variables was used to test each possible split for each node in each tree within the forest. 100 trees were grown and 60% of the sample was selected when performing the bagging process.

Number of Observations	
Type	N
Number of Observations Read	6337
Number of Observations Used	6337

Baseline Fit Statistics	
Statistic	Value
Average Square Error	0.215
Misclassification Rate	0.312
Log Loss	0.621

The number of observations used is equal to the number of observations read, which shows that there is no missing value in the response variable. I already set aside the missing value when I was cleaning the data.

The misclassification rate is as high as 31.2%. Only 68.8% of the sample is correctly classified.

Fit Statistics							
Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (OOB)	Misclassification Rate (Train)	Misclassification Rate (OOB)	Log Loss (Train)	Log Loss (OOB)
1	399	0.190	0.257	0.276	0.358	1.223	2.322
2	807	0.161	0.244	0.238	0.349	0.529	1.916
3	1154	0.156	0.234	0.226	0.338	0.490	1.538
4	1548	0.151	0.228	0.215	0.328	0.466	1.290
5	1934	0.149	0.225	0.210	0.327	0.461	1.136
6	2349	0.147	0.221	0.206	0.327	0.458	0.991
7	2760	0.147	0.220	0.206	0.324	0.456	0.916
8	3178	0.145	0.217	0.206	0.321	0.454	0.818
9	3576	0.145	0.215	0.204	0.317	0.453	0.779
10	3979	0.144	0.212	0.204	0.316	0.452	0.716
11	4382	0.144	0.211	0.199	0.314	0.450	0.690

90	36421	0.139	0.197	0.195	0.298	0.439	0.581
91	36812	0.139	0.197	0.194	0.298	0.439	0.581
92	37201	0.139	0.197	0.194	0.298	0.439	0.581
93	37579	0.139	0.197	0.193	0.297	0.439	0.581
94	37959	0.139	0.197	0.193	0.298	0.439	0.581
95	38355	0.139	0.197	0.193	0.299	0.439	0.581
96	38761	0.139	0.197	0.193	0.299	0.439	0.582
97	39135	0.139	0.197	0.193	0.299	0.439	0.581
98	39506	0.139	0.197	0.194	0.299	0.439	0.581
99	39881	0.139	0.197	0.194	0.299	0.439	0.581
100	40264	0.139	0.197	0.193	0.299	0.439	0.581

This is part of the fit statistics. In total, 100 trees were grown.

Loss Reduction Variable Importance					
Variable	Number of Rules	Gini	OOB Gini	Margin	OOB Margin
H1GH6	2215	0.012827	0.00462	0.025655	0.018496
H1FS11	2912	0.012193	0.00274	0.024386	0.015455
BIO_SEX	2320	0.005163	-0.00095	0.010326	0.003832
H1GH2	3959	0.009306	-0.00439	0.018612	0.005786
H1WP9	3100	0.008223	-0.00489	0.016447	0.004571
H1WP13	3144	0.009267	-0.00508	0.018535	0.004702
H1GI20	4728	0.012520	-0.01310	0.025039	0.001247
BMI	17786	0.077537	-0.04255	0.155075	0.035034

This shows the importance of variables in predicting the response variable. We can see the *H1GH6*, *H1FS11* and *BIO_SEX* are the 3 variables with most importance.

My code:

```

1 *Load data;
2 LIBNAME mydata "/courses/d1406ae5ba27fe300" access=readonly;
3 data new; set mydata.addhealth_pds;
4
5 if H1GI20=97 then delete; if H1GI20=99 then delete; if H1GI20=96 then delete;
6 if H1GI20=98 then delete;
7 if H1GH1=6 then H1GH1=.; if H1GH1=8 then H1GH1=.;
8 if H1GH2=6 then H1GH2=.; if H1GH2=8 then H1GH2=.;
9 if H1GH6=6 then H1GH6=.; if H1GH6=8 then H1GH6=.;
10 if H1GH59A=96 then H1GH59A=.; if H1GH59A=98 then H1GH59A=.; if H1GH59A=99 then H1GH59A=.;
11 if H1GH59B=96 then H1GH59B=.; if H1GH59B=98 then H1GH59B=.; if H1GH59B=99 then H1GH59B=.;
12 if H1GH60=996 then H1GH60=.; if H1GH60=998 then H1GH60=.; if H1GH60=999 then H1GH60=.;
13 if H1FS11=6 then H1FS11=.; if H1FS11=8 then H1FS11=.;
14 if H1WP9=6 then H1WP9=.; if H1WP9=7 then H1WP9=.;
15 if H1WP9=8 then H1WP9=.; if H1WP9=9 then H1WP9=.;
16 if H1WP13=6 then H1WP13=.; if H1WP13=7 then H1WP13=.;
17 if H1WP13=8 then H1WP13=.; if H1WP13=9 then H1WP13=.;
18
19 H1GH59=H1GH59A * 12 + H1GH59B; /*add a new variable*/
20 BMI=H1GH60 * 0.454 / (H1GH59 * 0.0254)**2; /*body mass index*/
21
22 if H1GH1 <= 2 then health = 1;
23 else health = 0;
24
25 label AID="respondent ID"
26       BIO_SEX="gender"
27       H1GI20="grade"
28
29       H1GH1="general health"
30       H1GH2="frequency of headache"
31       H1GH6="frequency of feeling weak"
32       H1GH59A="height in feet"
33       H1GH59B="height in inch"
34       H1GH60="weight (pound)"
35       H1FS11="feeling happy"
36       H1WP9="how close with mother"
37       H1WP13="how close with father"
38       H1GH59="height (inch)";
39
40
41
42 *Run the random forest;
43 proc hpforest;
44 target health/level=nominal;
45 input BIO_SEX H1GI20 H1GH2 H1GH6 H1FS11 H1WP9 H1WP13/level=nominal;
46 input BMI / level=interval;
47 run;

```