Respiratory Illness Prediction

Zhirou Zhou

Dec 11, 2019

Overview

1. Preprocessing
   - Variable Transformation
   - Missing Values
   - Feature Selection
2. Validation
3. Classification Tree
4. Random Forest

Variable Transformation

- Numeric: normalization
- Sequential questions

Any furry animals	How many pets
No	NA
Yes	3

Pets
0
3

- Virus and bacteria testing: ‘tlda’

Alias	visit_id	Adeno	B.pert
C5E05	2	FALSE	FALSE
C5E05	4	TRUE	FALSE

- Vaccination record: ‘vaccines’

Alias	pCGA	DTaP (2 months)	DTaP (4 months)
C01D8	44	0	0
C01D8	52	1	0
C01D8	61	1	1

Missing Values

- One row per subject
  - Mean
  - Most frequent category
- Multiple records per subject
  - Last observation carry forward/backward
  - Vaccines: most frequent category grouping by pCGA (whether most of other babies with same pCGA received that dose of vaccine)

Feature Selection

- Correlation

	Weight	HC	Length	Age	Temp	BMI
Weight	1	0.938	0.947	0.910	0.040	-0.025
HC		1	0.944	0.920	0.057	-0.036
Length			1	0.927	0.048	-0.068
Age				1	0.041	-0.035
Temp					1	-0.013
BMI						1

Feature Selection

- Collapse infrequent classes

#Pets	#Obs
0	2020
1	155
2	97
3	43
4	11
5	2
6	2
8	2
10	6
28	1

#Pets	#Obs
0	2020
1	155
2	97
3	43
4	11
5	13

Validation

- 10-fold cross validation
- Blocked by Alias
- Over-sampling to balance the target variable

Tuning Parameter

- Grow a large tree T_0 by stopping the splitting process when some minimum node size is reached
- Prune T_0 using *cost-complexity pruning*
  - Define a subtree $T \subset T_0$
  - t is the number of terminal nodes in T

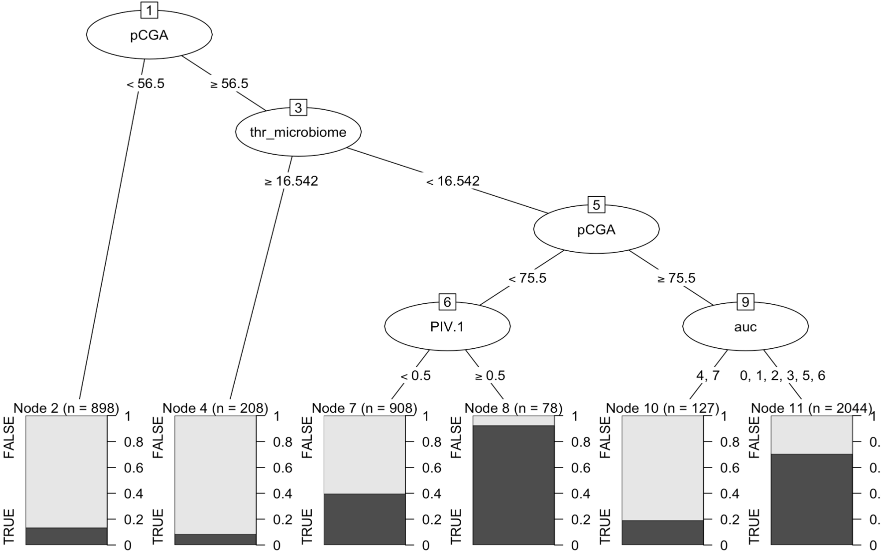
$$N_m = \#\{x_i \in R_m\}\ ,\ \ \hat{c}_m\ ,\ \ Q_m(T) \text{ (Node Impurity)},$$

  - Define the cost complexity criterion

$$C_\alpha(T) = \sum_{m=1}^{t} N_m Q_m(T) + \alpha t$$

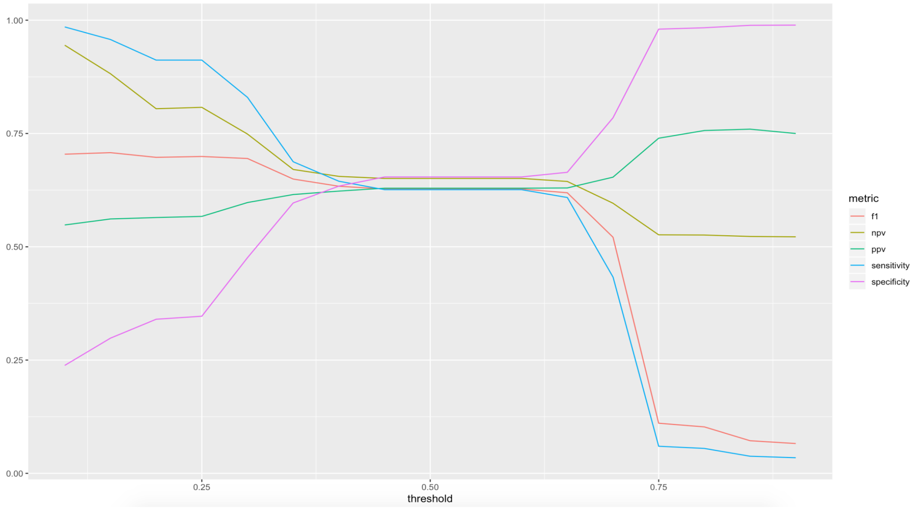
  - For each α, there is a unique smallest subtree T_α that minimizes $C_\alpha(T)$

Results: is_illness



minimum node size $= 50$, $\alpha = 0.03$

Results: threshold

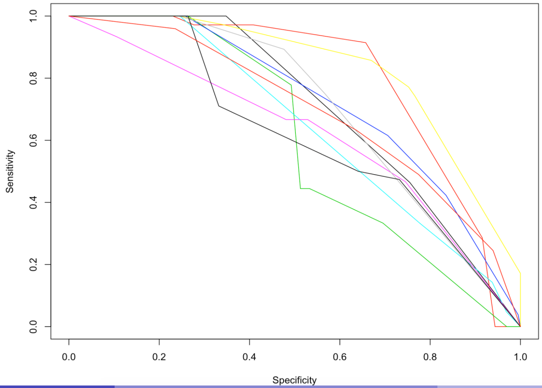


threshold = 0.5

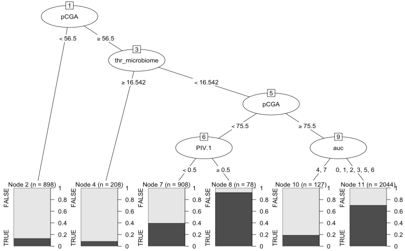
Results

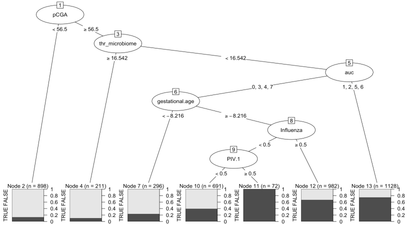
	is_illness	
Prediction	FALSE	TRUE
FALSE	1458	782
TRUE	771	1309

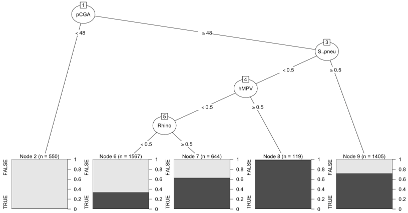
Accuracy	0.6405
PPV	0.6293
NPV	0.6509
F1	0.6277
AUC	0.7034

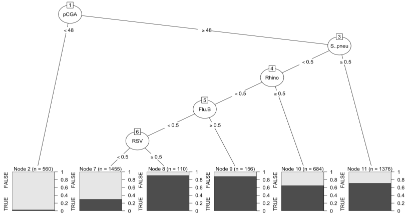


Variation


1
pCGA
< 56.5
≥ 56.5
3
thr_microbiome
≥ 16.542
< 16.542
5
pCGA
< 75.5
≥ 75.5
6
PIV.1
< 0.5
≥ 0.5
9
auc
4, 7
0, 1, 2, 3, 5, 6
Node 2 (n = 898)
Node 4 (n = 208)
Node 7 (n = 908)
Node 8 (n = 78)
Node 10 (n = 127)
Node 11 (n = 2044)
FALSE
TRUE
1
0.8
0.6
0.4
0.2
0


1
pCGA
< 56.5
≥ 56.5
3
thr_microbiome
≥ 16.542
< 16.542
5
auc
0, 3, 4, 7
1, 2, 5, 6
6
gestational.age
< −8.216
≥ −8.216
8
Influenza
< 0.5
≥ 0.5
9
PIV.1
< 0.5
≥ 0.5
Node 2 (n = 898)
Node 4 (n = 211)
Node 7 (n = 296)
Node 10 (n = 691)
Node 11 (n = 72)
Node 12 (n = 982)
Node 13 (n = 1128)
TRUE FALSE
1
0.8
0.6
0.4
0.2
0


1
pCGA
< 48
≥ 48
3
S..pneu
< 0.5
≥ 0.5
4
hMPV
< 0.5
≥ 0.5
5
Rhino
< 0.5
≥ 0.5
Node 2 (n = 550)
Node 6 (n = 1567)
Node 7 (n = 644)
Node 8 (n = 119)
Node 9 (n = 1405)
FALSE
TRUE
1
0.8
0.6
0.4
0.2
0


1
pCGA
< 48
≥ 48
3
S..pneu
< 0.5
≥ 0.5
4
Rhino
< 0.5
≥ 0.5
5
Flu.B
< 0.5
≥ 0.5
6
RSV
< 0.5
≥ 0.5
Node 2 (n = 560)
Node 7 (n = 1455)
Node 8 (n = 110)
Node 9 (n = 156)
Node 10 (n = 684)
Node 11 (n = 1376)
FALSE
TRUE
1
0.8
0.6
0.4
0.2
0

Random Forest

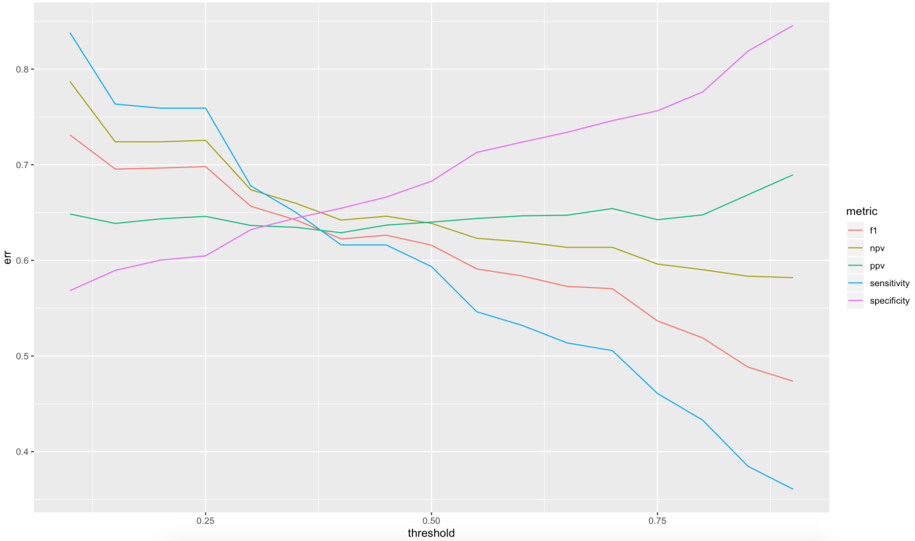
For $b = 1$ to B:

- Draw a bootstrap sample $\boldsymbol{Z}$ of size N from the training data
- Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached
    - Select m variables at random from the p variables
    - Pick the best variable/split-point among the m
    - Split the node into two child nodes
- Output the ensemble of trees $\{T_b\}_1^B$

- Classification prediction: $\hat{C}_b(x) =$ the prediction of the bth random-forest tree, then

$$\hat{C}_{rf}^B(x) = majority\ vote\{\hat{C}_b(x)\}_1^B$$

Results

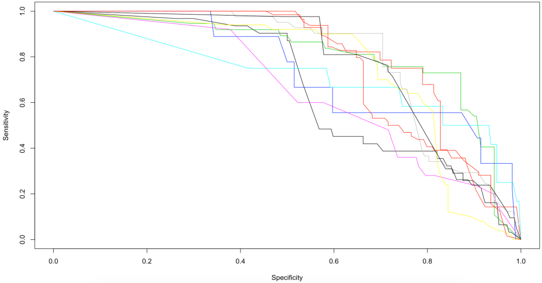


threshold = 0.375

Results

	is_illness	
Prediction	FALSE	TRUE
FALSE	1609	828
TRUE	620	1280

	Tree	Forest
Accuracy	0.6405	0.6661
PPV	0.6293	0.6737
NPV	0.6509	0.6602
F1	0.6277	0.6387
AUC	0.7034	0.7352



Q&A