Data Science - Practice 8 (Decision Tree)

Make sure you not only just "write down" the R code but also "explain the answer with your own language". All answers without explanation will not be accepted.

Problem

Import PRSA_set.csv. When running decision tree model, set the random number generator to 2022.

year	Year
month	Month
day	Day
hour	0h ~ 23h
pm2.5	Fine dust concentration (ug/mˆ3)
DEWP	Dew Point
TEMP	Temperature
PRES	Air pressure (hPa)
cbwd	Wind Direction
Iws	Cumulated wind speed (m/s)
Is	Snowfall per hour
Ir	precipitation per hour

< Question 1 – PRSA data >

(1) Remove missing pm2.5 from the data set. (2) Create a new variable called "bad_air", which returns TRUE if pm2.5 exceeds 75 and FALSE otherwise. (3) Create a new data set called "PRSA.train", which only contains the records between 2010 and 2013, and "PRSA.test", which only contains the records in 2014. (4) For this case, we did not use random sampling. Do you think it is acceptable? Explain.

< Question 2 – PRSA data >

Use all variables except 'time-related' variables to create a decision tree model that predicts "bad_air". Here, use the default setting of rpart() function. Measure accuracy, precision and recall, and F1 score. (Hereafter, let's call this baseline model.)

< Question 3 – PRSA data >

Test your model with 'PRSA.test' set. Is there an issue of "overfitting"? Explain why you think so.

< Question 4 – PRSA data >

(pre-pruning approach) Predict the same model with the maximum depth value of 3. How many decision tree scenarios are made from this? Explain each scenario with decision tree plot obtained from rplot.plot() function. (numbers presented in the figure should be explained.)

< Question 5 – PRSA data >

(pre-pruning approach) Predict the same model with the minimum number of observations of 15,000. Can this approach fix the issue of "overfitting"? State whether overfitting has been resolved and how the accuracy has been changed. Explain why this happened.

< Question 6 – PRSA data >

(post-pruning approach) Use post-prunning approach to predict your model. Explain the procedure and how you set the post-prunning values.

< Question 7 – PRSA data >

Draw ROC curve and measure AUC for models you created in Question 2 and 6 (Set type ='prob' in predict() to obtain estimated probability). Explain the results.