Data Science - Practice 7 (Single Var)

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_data_class.RData’. This data is hourly PM2.5 (fine dust concentration) records in Beijing.

No	row number
Month	year of data in this row
TEMP	Temperature
Iws	Cumulated wind speed (m/s)
time	‘day’ or ‘night’
pm2.5	‘HIGH’ if PM2.5 is higher than 75, ‘LOW’ otherwise.
type	‘train’ or ‘test’

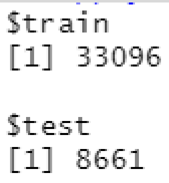
< Question 1 – PRSA data >

In this practice, we want to build a single variable model that uses Month to predict fine dust concentration. For this purpose, month and pm2.5 are the key variables. In this respect, create a new data set called “PRSA_DATA” by excluding the missing values from PRSA_data_class. Here, check out the missing data and explain whether just removing missing value is fine or not.

< Question 2 – PRSA data >

Using PRSA_DATA and type variable, create a list called ‘PRSA_data.ls’. Here, the first list element only contains train data type and the second list element contains the test data type. For this operation, use for iteration. Find the number of samples for each data set as shown below.

Expected Result



```
$train
[1] 33096

$test
[1] 8661
```

< Question 3 – PRSA data >

Using ‘PRSA_data.ls’, create a single variable model the predicts fine dust concentration using Month variable with threshold of 0.5 (Here, our interest is the case where pm2.5 is “High”). Find the accuracy of the model with train and test set. For this operation, follow the steps that you learned from the class and write your own explanation for each step.

Sample Result

```
[1] "Accuracy of train set: 0.537"
[1] "Accuracy of test set: 0.51"
```

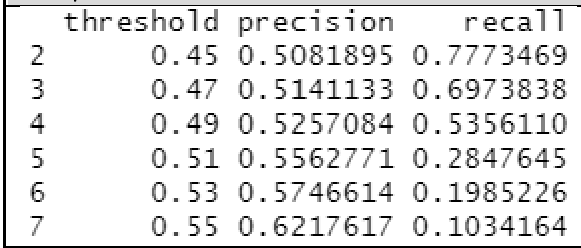
< Question 4 – PRSA data >

Calculate the AUC of our model for each train and test set. Also, draw a ROC curve. From this, what can we tell about our model?

< Question 5 – PRSA data >

Change the threshold as shown below, and find how the precision and recall are changing. Here, use train set. Explain why these values are changing and meaning of it with your own language, from “this data sample’s context (what is positive, false positive, false negative)”.

Sample Result



```
  threshold precision    recall
2      0.45 0.5081895 0.7773469
3      0.47 0.5141133 0.6973838
4      0.49 0.5257084 0.5356110
5      0.51 0.5562771 0.2847645
6      0.53 0.5746614 0.1985226
7      0.55 0.6217617 0.1034164
```

< Question 6 – PRSA data >

We learned that there is a trade-off relationship between precision and recall. F1 score is harmonic mean of precision and recall that can be used to explain the prediction performance. Measure it with the following equation and explain the meaning of the measure.

$$\text{F1Score} = 2 * \frac{precision * recall}{precision + recall}$$

< Question 7 – PRSA data >

Using ‘PRSA_data.ls’, create a single variable model the predicts fine dust concentration using TEMP variable. Repeat the operation you did in Question 3, 4, & 5.

< Question 8 – PRSA data >

Do you think either ‘Month’ or ‘TEMP’ single variable model is good enough? If you think one of them is good enough, explain why. If you think they are not good enough, suggest the “new” possible variable that you think that it will be useful to predict fine dust concentration.