

Data Science

Homework 06

1. 한동 honor code 에 맞게 과제를 진행하여 주세요.
 2. 과제의 경우 팀당 1 개의 결과물을 제출하면 됩니다.
 3. 과제 제출 기한은 ~5/2 23:59 입니다. (1 분당 0.1 감점) 시험기간이라 2 주로 늘렸습니다.
 4. 제출은 LMS>과제 및 평가>Homework06 로 하시면 됩니다. (팀 내 1 명이 제출)
 5. LMS 제출이 안되는 경우는 TA 이메일로 제출하시기 바랍니다. (22100733@handong.ac.kr)
-
1. Please proceed with the assignment following the Handong honor code.
 2. For assignments, one submission per team is sufficient.
 3. The deadline for assignment submission is until 4/25, 23:59. (0.1 points deducted per minute late)
 4. Submissions should be made to LMS>Assignments>Homework06. (One member of the team should submit)
 5. If you cannot submit via LMS, please submit to the TA email. (22100733@handong.ac.kr)

모든 학생들은 아래의 링크에 접속하여, 본 과제물에 대한 Peer Evaluation을 진행하여 주시기 바랍니다.
제출시간 마감 이전에 응답하지 않으면 불이익이 있을 수 있습니다.

All students are requested to access the link below and conduct a Peer Evaluation for this assignment.

01분반(KOR) - <https://forms.gle/zLKhpja1UKnpXW6T6>
2rd Class(ENG) - <https://forms.gle/8wHcAzk7RK9dvbaX8>

Data Science - Practice 6

모든 문제에 대하여 코드만 작성하지 말고 데이터를 해석한 결과를 함께 작성하시오.

Put your explanation of your findings from dataset with your answer in your report as well as R code.

loading data into R

아래와 같이 실습을 위한 데이터를 R에 loading 하시오.

Load the dataset by typing in code below.

```
PRSA_data <- load("PRSA_data.RData")
```

Data description

```
## # A tibble:      11 x 6
##       No Month  TEMP    lws  time pm2.5
##   <dbl> <fct>   <dbl> <dbl> <fct> <fct>
## 1     1     1 Jan    -11  1.79 night NA
## 2     2     2 Jan    -12  4.92 night NA
## 3     3     3 Jan    -11  6.71 night NA
## 4     4     4 Jan    -14  9.84 night NA
## 5     5     5 Jan    -12 13.00 night NA
## 6     6     6 Jan    -10 16.10 night NA
## 7     7     7 Jan     -9 19.20 night NA
## 8     8     8 Jan     -9 21.00 night NA
## 9     9     9 Jan     -9 24.20 night NA
## 10    10    10 Jan     -8 27.30 night NA
## 11    11    11 Jan     -7 31.30 day   NA
```

variable	의미	description
No	인덱스	Index
Month	월	Month
TEMP	기온	Temperature
lws	누적풍속	Cumulated wind speed (m/s)
time	측정시간	Time of Measurement
pm2.5	미세먼지농도(75이상 High, 이하 Low)	Fine dust concentration (over 75 = high, lower 75 = Low)

< Question 1 -1>

미세먼지 농도(pm2.5)를 예측하는 Single Variable 모델을 만들어보려고 한다. 모델을 만들기 전, 데이터 탐색을 진행하려 한다. 가장 먼저 train과 test 데이터의 타입을 확인해보자

Sample Result

```
## Classes      'tbl_df', 'tbl' and 'data.frame':      35064 obs. of  6 variables:
## $ No         : num      1  2  3  4  5  6  7  8  9 10 ...
## $ Month      : Factor w/ 12 levels "Apr", "Aug", "Dec",... 5  5  5  5  5  5  5  5 ...
## $ TEMP       : num     -11 -12 -11 -14 -12 -10 -9 -9 -9 -8 ...
## $ lws        : num      1.79 4.92 6.71 9.84 12.97 ...
## $ time       : Factor w/ 2 levels "day", "night": 2  2  2  2  2  2  2  2  2  2 ...
## $ pm2.5      : Factor w/ 2 levels "HIGH", "LOW": NA NA NA NA NA NA NA NA NA ...

## Classes      'tbl_df', 'tbl' and 'data.frame':      8760 obs. of  6 variables:
## $ No         : num    35065 35066 35067 35068 35069 ...
## $ Month      : Factor w/ 12 levels "Apr", "Aug", "Dec",... 5  5  5  5  5  5  5  5 ...
## $ TEMP       : num       7  7  6  6  3  4  6  6  6  7
## $ lws        : num    143.48 147.5 151.52 153.31 0.89 ...
## $ time       : Factor w/ 2 levels "day", "night": 2  2  2  2  2  2  2  2  2  2 ...
## $ pm2.5      : Factor w/ 2 levels "HIGH", "LOW": 2  2  2  2  1  1  1  1  2  2 ...
```

< Question 1 -2>

분석 과정에서 결측치를 살펴보고 처리하는 것은 매우 중요한 과정이다. train과 test에 결측치가 얼마나 있는지 열 별로 확인해보자.

Sample Result

```
## No    Month  TEMP    lws    time    pm2.5
##    0         0      0      0      0      1968

## No    Month  TEMP    lws    time    pm2.5
##    0         0      0      0      0      99
```

< Question 1 -3>

결측치가 주로 몰려있는 월이나 시간이 있는지 확인해보자.

Sample Result

```
## Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep
## 192 372 151 8 242 66 178 163 109 85 140 262

## Apr Aug Dec Feb Jan Jul Jun Mar May Nov Oct Sep
## 3 11 28 3 7 7 17 1 2 14 1 5
```

```
## day night
## 744 1224

## day night
## 43 56
```

< Question 1 -4>

pm2.5는 목적변수이므로 NA가 허용되지 않는다. 이를 삭제해보자.

Sample Result

```
## No Month TEMP lws time pm2.5
## 0 0 0 0 0 0

## No Month TEMP lws time pm2.5
## 0 0 0 0 0 0
```

< Question 2-1 >

Month 변수를 활용하여 pm2.5를 예측하는 Single variable 모델을 만들어보자.

이때 Threshold는 0.5로 설정하도록 하자.

또한 이 모델의 정확도(Accuracy)를 train과 test 데이터에서 계산해보도록 하자.

- 일반적으로 예측 모형에서 예측 대상이 되는 (관심대상인) class의 sample을 positive sample로 간주합니다.

여기에서는 미세먼지가 나쁨인 pm2.5 = HIGH인 경우를 positive(TRUE) sample로 보고 문제를 풀어봅시다.

일반적인 convention을 꼭 따르지 않아도 모델링에 달라지는 것은 없지만 communication을 잘하기 위해서는 convention을 따르는 것이 편리합니다.

Let's train a single variable model that predicts **pm2.5** using **Month** variables.

Set the threshold to 0.5 at this time.

Find the accuracy of this model with train and test data.

- we typically consider the samples of class we are interested in for prediction or estimation as positive samples

So here we consider the day of pm2.5 = HIGH where air is in bad condition because of fine dust concentration.

You do not have to follow the convention always but for good communication, we'd better follow the convention to avoid the chance of miscommunication.

Sample Result

```
## [1] "accuracy for train data 0.54"  
## [2] "accuracy for test data 0.51"
```

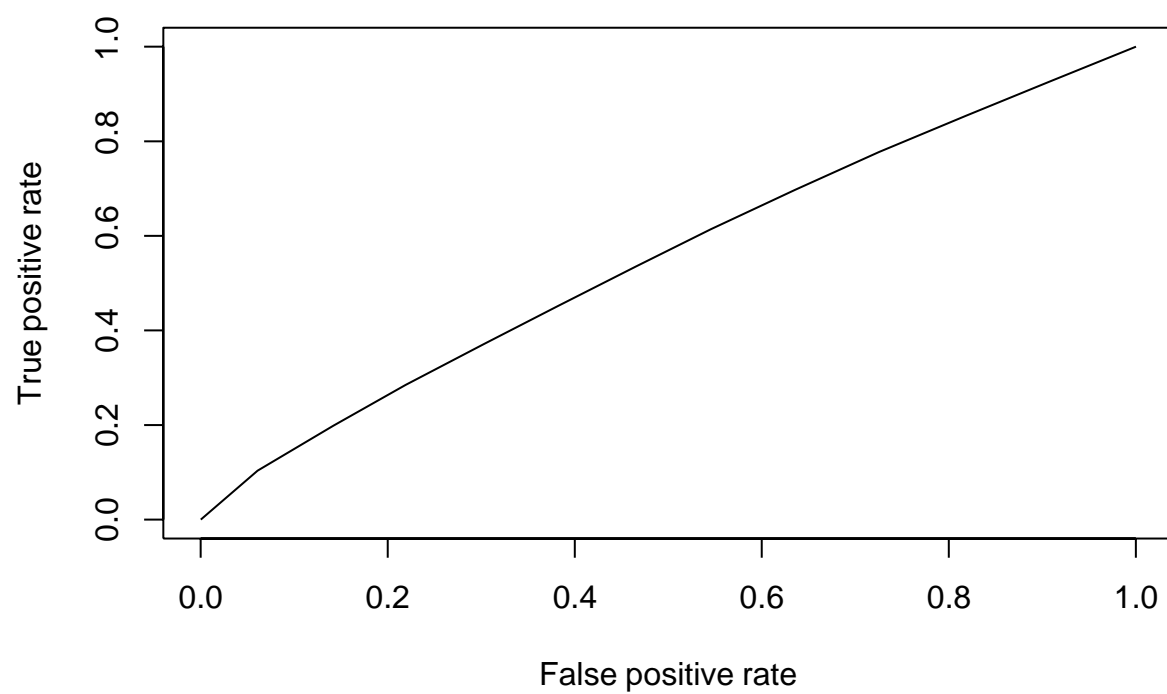
< Question 2-2 >

문제 2-1에서 구한 모델의 AUC를 train과 test 데이터 각각에 대해서 계산해보고, ROC 커브를 그려보자.

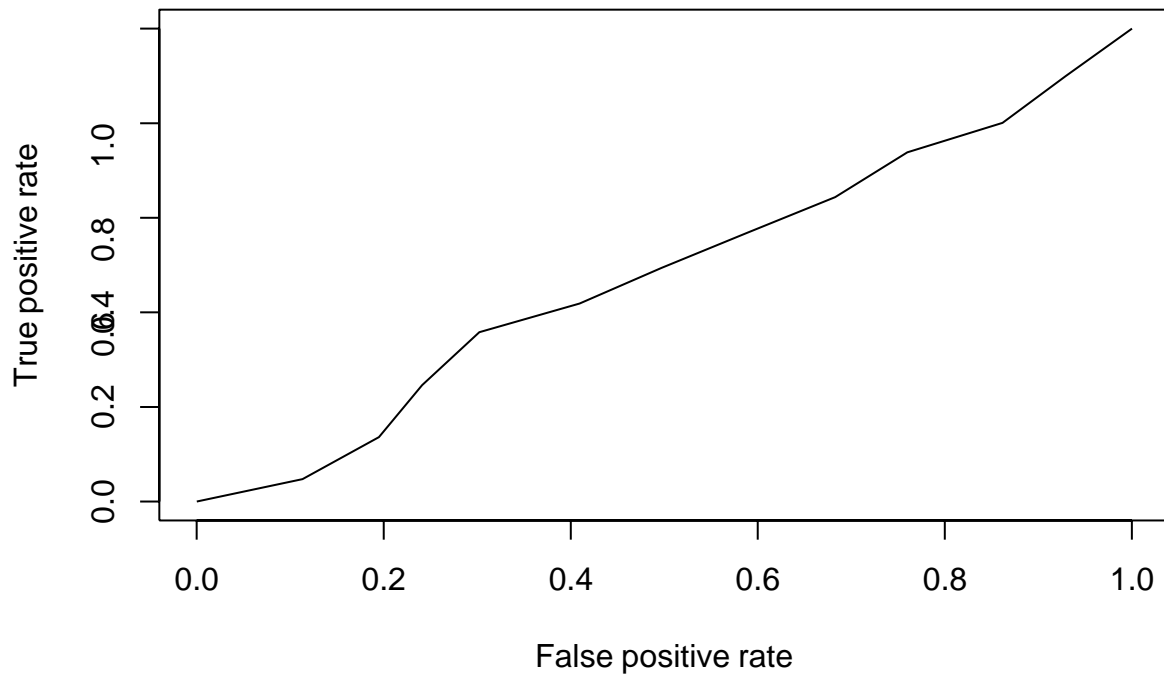
Calculate the AUC of the model from question 2-1 for each of the train test data and draw a ROC curve.

Sample Result

```
## [1] "AUC value for train data 0.55"
```



```
## [1] "AUC value for test data 0.48"
```



< Question 2-3 >

문제 2-1, 2-2를 바탕으로 이 모델은 과적합(overfitting)인지, 아닌지 설명하여보라.

Based on questions 2-1 and 2-2, explain whether this model is overfitting.

< Question 2-4 >

문제 2-1에서 구한 모델의 threshold를 바꿔가면서 precision과 recall 값의 변화를 확인해보자.

As you change the threshold of the model from question 2-1, find how the precision and recall are changing.

Sample Result

##	threshold	precision	recall
## 1	0.45	0.5081895	0.7773469
## 2	0.47	0.5141133	0.6973838
## 3	0.49	0.5257084	0.5356110
## 4	0.51	0.5562771	0.2847645
## 5	0.53	0.5746614	0.1985226
## 6	0.55	0.6217617	0.1034164

< Question 2-5 >

문제 2-4에서 나온 결과를 바탕으로 threshold를 어떻게 설정하는 것이 좋을지 생각해보고, 그 이유와 함께 설명해보자.

What would be the best threshold setting for the model for what you know from Q2-4?

What makes you think so?

< Question 2-6 >

Trade-off 관계에 있는 precision과 recall을 하나의 measure로 보기 위해서 F1 Score라는 것을 사용하기도 한다.

문제 2-4에서 선정한 threshold를 기준으로 F1 score를 계산해보자. F1 score는 아래 식과 같이 계산한다.

As precision and recall are the measures in trade-off, F1 score is harmonic mean of precision and recall that can be used to explain the prediction performance.

Let's find the F1 score based on the best threshold setting you have found in question 2-4. The formula for F1 score is shown below.

$$F1\ score = 2 \times \frac{precision+recall}{precision+recall}$$

```
## [1] "F1 score for test data 0.66"
```

< Question 3 >

TEMP 변수를 사용해서 문제 2번의 과정을 반복해보자.

hint) TEMP 변수를 그룹화하여 모델의 성능을 높일 수 있다. 적절한 그룹화 기준을 고려해보자.

Repeat the tasks in question 2 using the **TEMP** variable.