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 type in code below.

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

Data description

```
## # A tibble: 11 x 6
##       No Month  TEMP   Iws time  pm2.5
##    <dbl> <fct> <dbl> <dbl> <fct> <fct>
##  1     1 Jan     -11  1.79 night NA
##  2     2 Jan     -12  4.92 night NA
##  3     3 Jan     -11  6.71 night NA
##  4     4 Jan     -14  9.84 night NA
##  5     5 Jan     -12 13.0  night NA
##  6     6 Jan     -10 16.1  night NA
##  7     7 Jan      -9 19.2  night NA
##  8     8 Jan      -9 21.0  night NA
##  9     9 Jan      -9 24.2  night NA
## 10    10 Jan      -8 27.3  night NA
## 11    11 Jan      -7 31.3  day   NA
```

variable	의미	description
No	인덱스	Index
Month	월	Month
TEMP	기온	Temperature
Iws	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 >

미세먼지 농도(pm2.5)를 예측하는 Single Variable 모델을 만들어보려고 한다.

가장 먼저 train과 test 데이터의 타입을 확인해보자. 또한 train과 test 데이터에서 NA가 얼마나 있는지 확인하고, 주로 몰려있는 날짜나 기간이 있는지 확인하여보자.

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

We are building a single variable model that predicts fine dust concentration (pm2.5).

First, let's check the types of variables in train and test dataset. And check how many missing values(NAs) are in the train and test data, and check whether missing values have occurred and been concentrated in some specific period of time such as season, month, or week or some other condition.

pm2.5 is the outcome variable to predict. So Missing value is not allowed for **pm2.5**.

Exclude observations that pm2.5 is not available (NA).

Sample Result

train dataset

```
## Rows: 35,064
## Columns: 6
## $ No    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 1~
## $ Month <fct> Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan,~
## $ TEMP  <dbl> -11, -12, -11, -14, -12, -10, -9, -9, -9, -8, -7, -5, -5, -3, -2~
## $ Iws   <dbl> 1.79, 4.92, 6.71, 9.84, 12.97, 16.10, 19.23, 21.02, 24.15, 27.28~
## $ time  <fct> night, night, night, night, night, night, night, night, night, n~
## $ pm2.5 <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

test dataset

```
## Rows: 8,760
## Columns: 6
## $ No    <dbl> 35065, 35066, 35067, 35068, 35069, 35070, 35071, 35072, 35073, 3~
## $ Month <fct> Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan,~
## $ TEMP  <dbl> 7, 7, 6, 6, 3, 4, 6, 6, 6, 7, 8, 9, 10, 11, 11, 11, 11, 10, 9, 9~
## $ Iws   <dbl> 143.48, 147.50, 151.52, 153.31, 0.89, 4.02, 8.94, 16.09, 21.90, ~
## $ time  <fct> night, night, night, night, night, night, night, night, night, n~
## $ pm2.5 <fct> LOW, LOW, LOW, LOW, HIGH, HIGH, HIGH, HIGH, LOW, LOW, LOW, LOW, ~
```

< Question 2-1 >

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

이때 Threshold는 0.5로 설정하도록 하자.

또한 이 모델의 정확도(Accuracy)을 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 miscommuncation

Sample Result

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

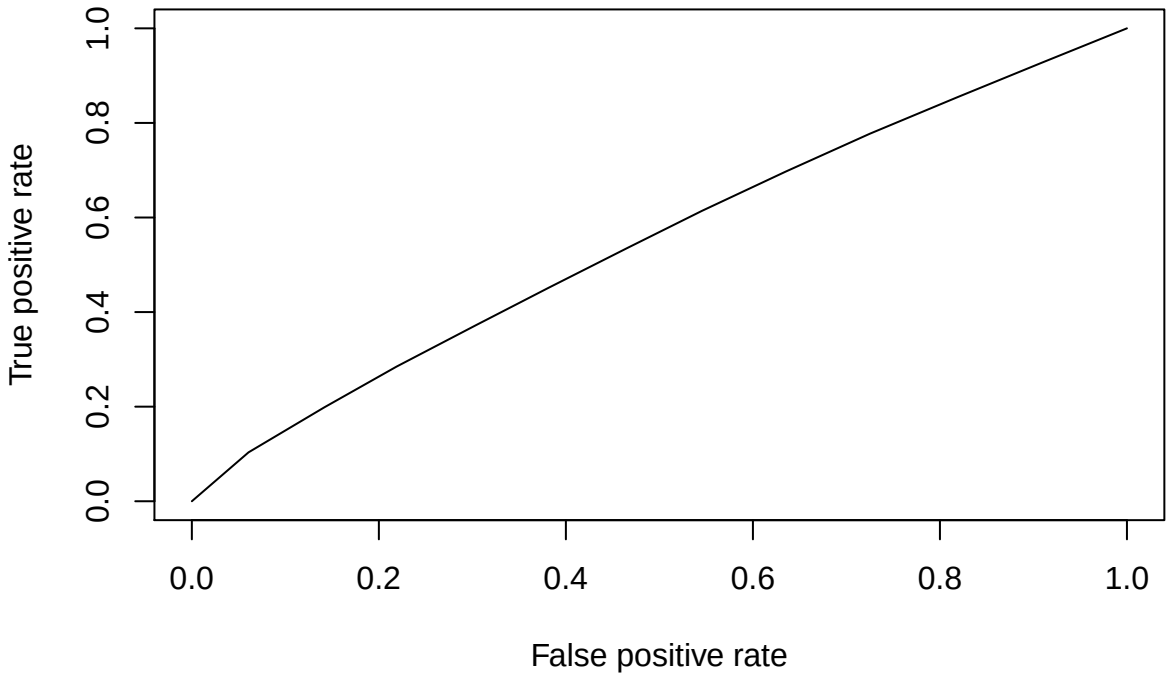
< Question 2-2 >

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

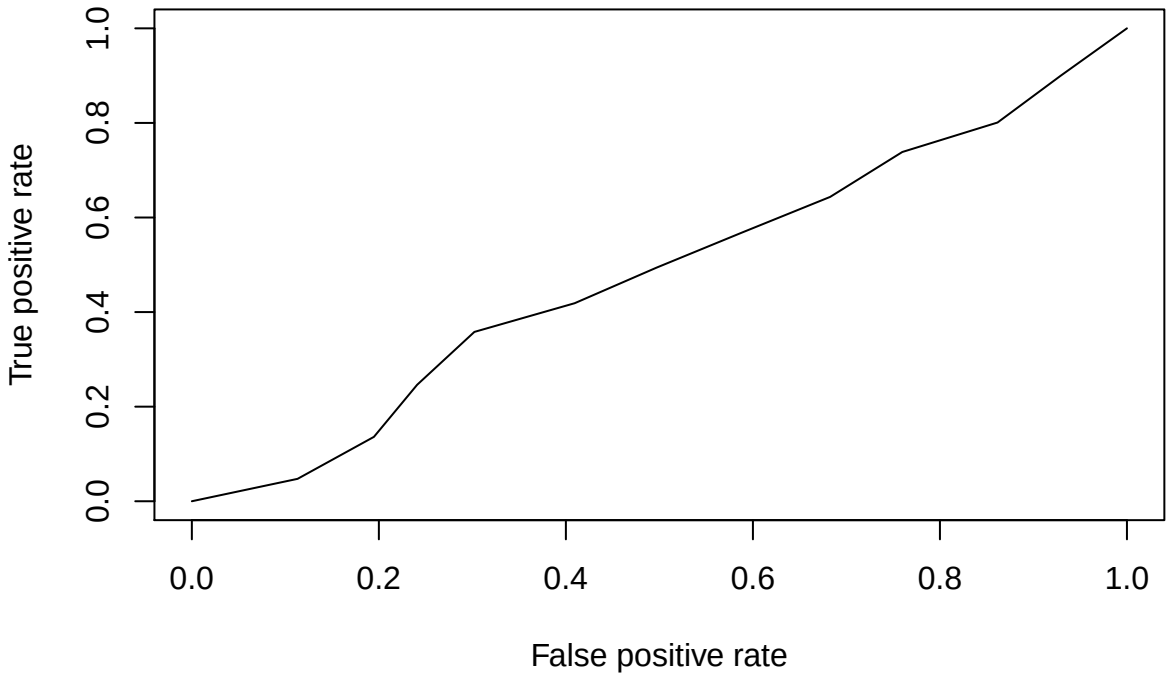
Calculate the AUC of the model from question 2-1 for each of the train, test data and draw a ROC curve.

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 meaures 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.

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

< Question 3 >

TEMP 변수를 사용해서 문제 2번의 과정을 반복해보자.

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

< Question 4 >

lws 변수를 사용해서 문제 2번의 과정을 반복해보자.

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

< Question 5 >

time 변수를 사용해서 문제 2번의 과정을 반복해보자.

Repeat the process in question 2 using the **time** variable.

< Question 6 >

2번에서 5번까지의 과정을 통해 얻은 모델들 중 어떤 변수를 사용했을 때 가장 예측 성능이 높은 모델이 만들어졌는지 확인하고, 성능이 높게 나온 이유가 무엇인지 추론하여 설명해보자.

From Q2 to Q5, what is the best performing prediction model for **pm2.5**?

And explain why the best model outperforms others.