<다변량분석 리포트 4>

2014170852 산업경영공학부 조영관

<Data Set>

미국 대학원 지원자들의 여러 가지 점수와 대학 등급에 따른 각 지원자들의 합격할 확률을 기록한 데이터

변수: Serial No, GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research, Chance of admit(target variable Y)

<모형 구축을 위해 필요하지 않은 변수 제거>

위의 변수들을 살펴보면, Serial No는 관측치의 번호를 나열한 것이므로 불필요한 변수이다. 따라서, Serial No를 제거해준다.

<입력 변수의 단변량 통계량 계산 및 BOX PLOT 그리기 & 정규분포를 따르는 변수 살펴보기>

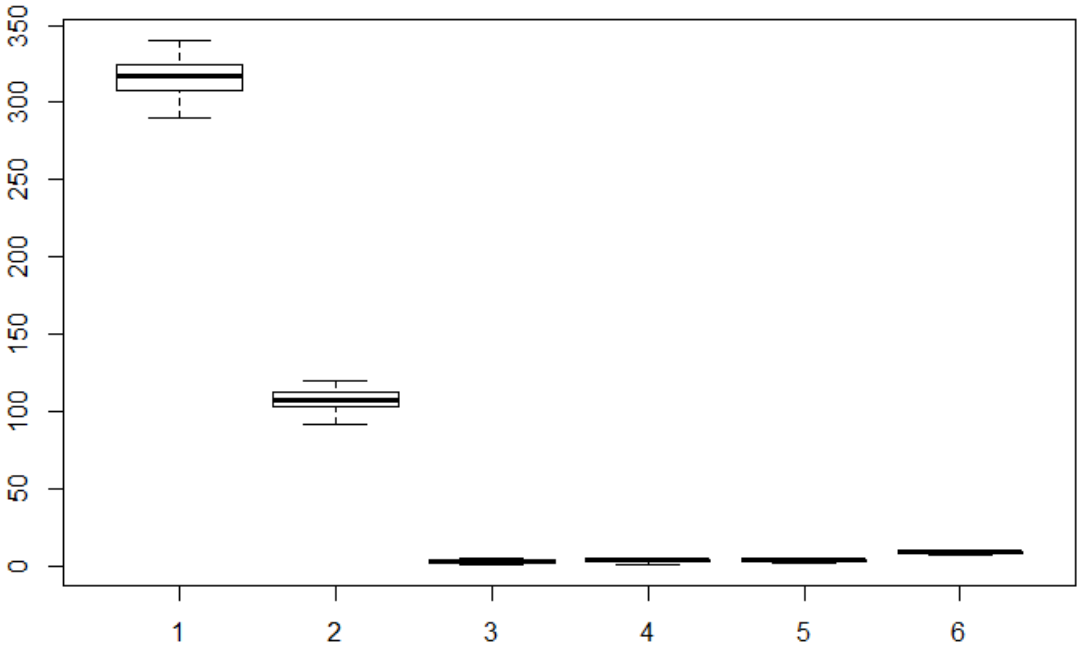
	mean	std	skewness	kurtosis
GRE.Score	316.9924242	11.3814169	-0.07098162	2.326104
TOEFL.Score	107.5176768	6.0010287	0.06812888	2.422757
University.Rating	3.1035354	1.1371239	0.16880534	2.200411
SOP	3.4078283	1.0030154	-0.26799041	2.300339
LOR	3.4671717	0.8881402	-0.08521422	2.285952
CGPA	8.6110354	0.5852899	-0.01104870	2.431140

각각의 입력변수에 대한 통계량 계산 값이다.

평균, 표준편차, skewness, kurtosis를 측정하였다.

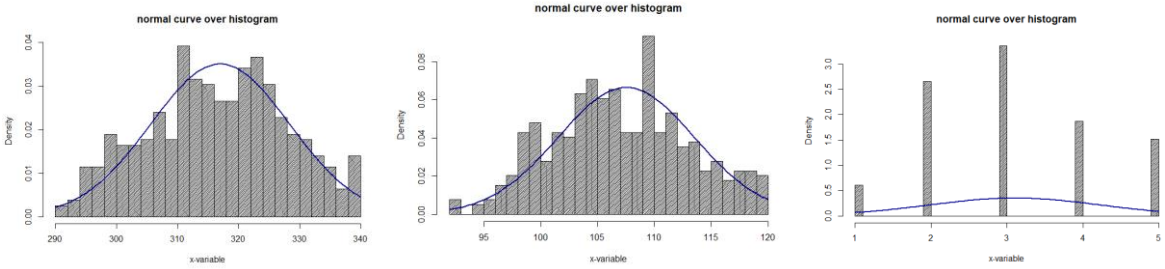
결과 값을 보면 6개의 변수들 모두 분산이 크지 않으며, skewness도 0에 근접하고 kurtosis도 정상에서 크게 벗어난 값이 없음을 확인할 수 있다. 6개의 변수들 모두 정규분포를 따를 것임을 추정할 수 있다. 실제로 그러한 결과를 보이는지 뒤에서 살펴보자.

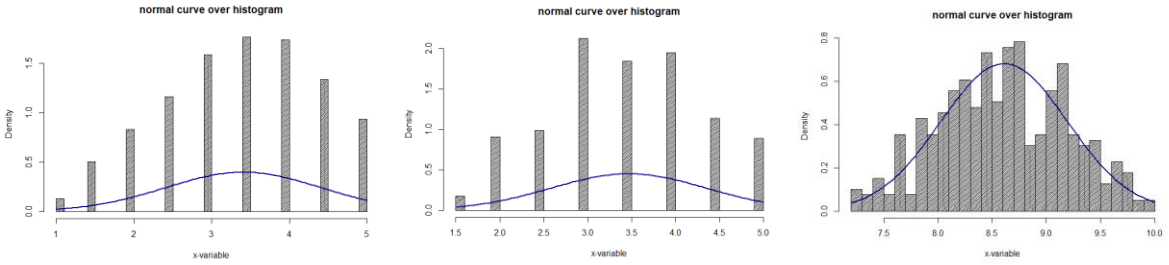
Box plot을 그려보자.



Box plot의 결과는 위와 같다. 왼쪽부터 순서대로 GRE.Score, TOEFL.Score, University.Rating, SOP, LOR, CGPA이다. 위 결과를 보면, 6개의 변수 모두 분산이 크지 않음을 알 수 있다.

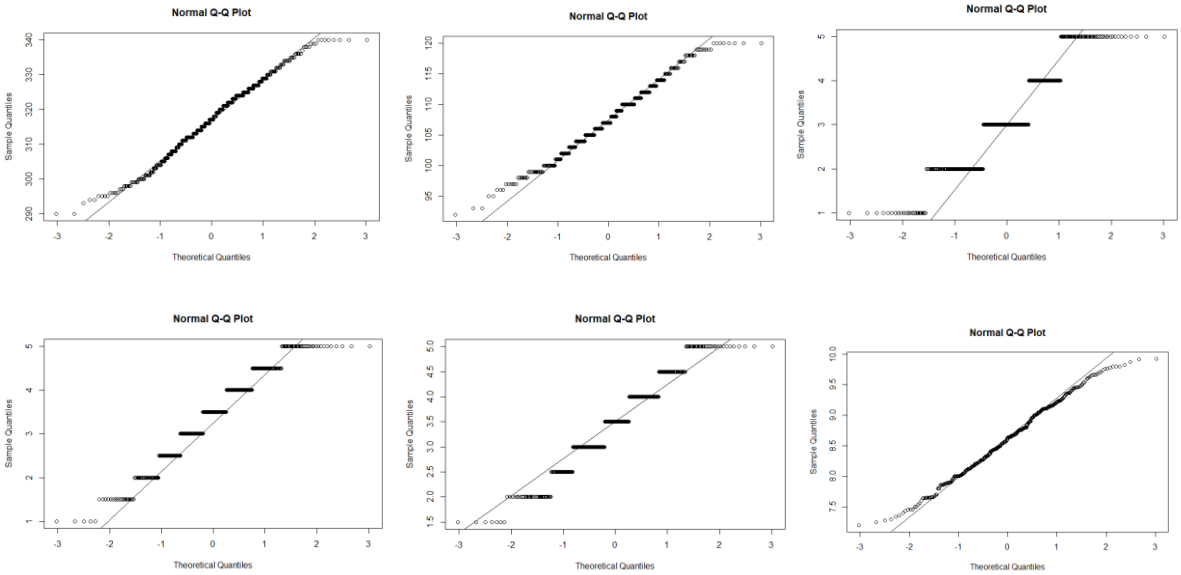
Histogram과 qqplot을 통해 정규분포를 따르는지 확인해보자.





위 왼쪽부터 순서대로 GRE.score, TOEFL.score, University.rating, SOP, LOR, CGPA이다.

University.rating, SOP, LOR의 경우 정확히는 연속형 값들은 아니고, 정해진 값들이 있기 때문에 위와 같은 모양이 나왔지만, 6개 변수 전체적으로 히스토그램을 보면 정규분포를 띄는 것으로 보인다.



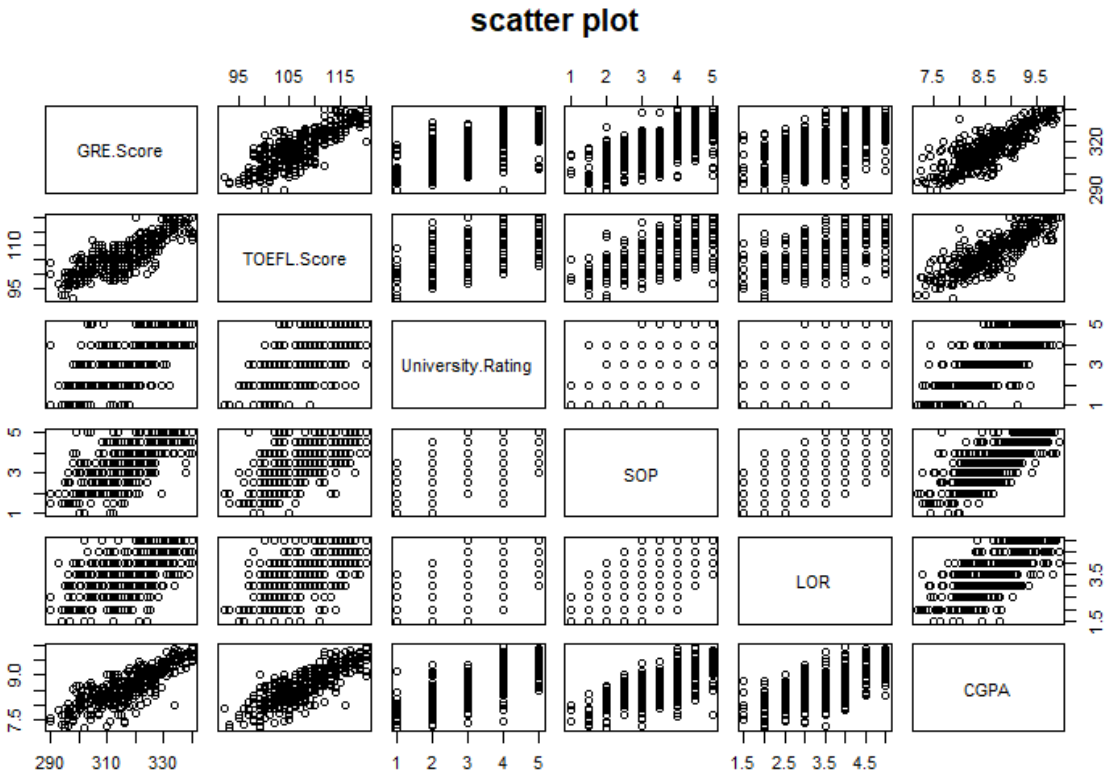
QQPLOT을 살펴보면 University.rating, SOP, LOR 위 세 가지는 히스토그램처럼, 연속형 값들이 아니라 정해진 값들이 있어서 가로로 일자 모양의 형태가 나온다. (그것을 감안하였을 때) 6개의 변수 모두 정규성을 가짐을 확인할 수 있다.

<BOX PLOT을 확인한 후 이상치 제거하기>

BOX PLOT을 모두 살펴보면 이상치 값들이 존재하지 않는다.

실제로, 이상치를 찾아 제거해주는 함수를 돌려본 결과, 위 데이터에는 이상치 값이 없었다.

<Scatter plot & Correlation plot 그리기, 변수들 간 상관관계 확인하기>

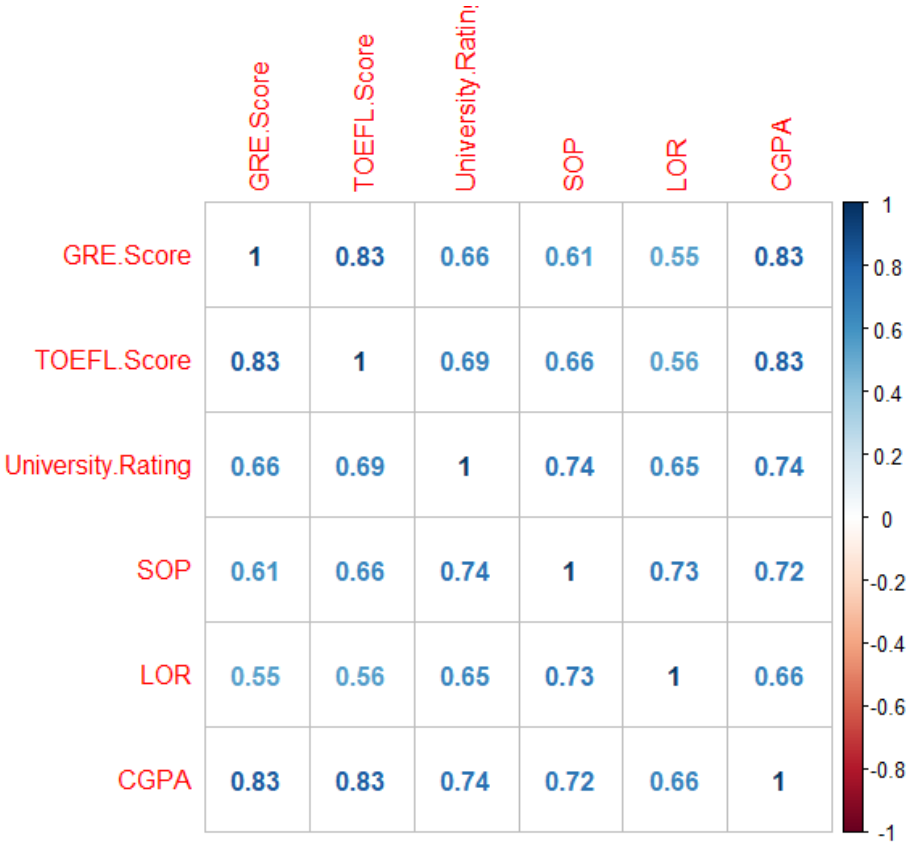


Scatter plot을 그려보면 위와 같은 결과가 나온다.

위 그림 결과를 보면, TOEFL.SCORE와 GRE.SCORE 간에는 양의 상관관계가 있음을 알 수 있다.

또한 GRE.SCORE와 CGPA도 양의 상관관계, TOEFL.SCORE와 CGPA도 양의 상관관계를 보임을 확인할 수 있다. 이 세 가지뿐 아니라 다른 변수들 간 관계도 이 셋만큼 뚜렷하지는 않지만 약한 양의 선형성을 띄고 있다.

다음으로 CORRPLOT을 살펴보자.



CORRPLOT의 결과는 위와 같다.

위 산점도에서 확인한대로, TOEFL.SCORE와 GRE.SCORE가 강한 상관관계, GRE.SCORE와 CGPA가 강한 상관관계, CGPA와 TOEFL.SCORE가 강한 상관관계를 보임을 확인할 수 있다.

그 외에 다른 변수들 사이에도 다 0.5 이상의 상관관계가 존재함을 확인할 수 있다.

즉, 6개의 변수들 모두 어느 정도 높은 상관관계가 존재한다.

<Target variable을 binary로 변환, Logistic regression 모델 구축, 유의미한 변수 확인하기>

Cut-off value를 0.8로 잡고 y값은 0 or 1 이진수로 변환하였다. 그 후 training set과 test set을 7대 3으로 나누었다. 그리고 Logistic regression 모델을 학습하였다.

그 결과 다음과 같았다.

```
Call:
glm(formula = Admit ~ ., family = "binomial", data = data_ynew_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0947  -0.1678  -0.0338   0.0816   3.2546

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.4997     0.6593  -5.308 1.11e-07 ***
GRE.Score           0.5842     0.7180   0.814 0.415876
TOEFL.Score         0.7710     0.5701   1.352 0.176282
University.Rating   0.7955     0.4311   1.845 0.065017 .
SOP                -0.3643     0.5543  -0.657 0.511090
LOR                 0.3687     0.4018   0.917 0.358910
CGPA                3.0734     0.8917   3.447 0.000567 ***
Research1           0.6935     0.6554   1.058 0.289993
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 337.23  on 277  degrees of freedom
Residual deviance:  99.69  on 270  degrees of freedom
AIC: 115.69

Number of Fisher Scoring iterations: 8
```

유의수준 0.1에서 위의 입력변수들의 p-value를 살펴보자.

Target variable Y값에 유의미한 영향을 미치는 변수는 University.Rating, CGPA 이렇게 두 가지가 있다. 다른 변수들은 모두 0.1보다 p-value 값이 커서 y에 유의미한 영향을 미치지 못한다.

<Confusion matrix, performance 지표 확인>

```
          lr_predicted
lr_target  0  1
        0 84  1
        1 12 23
```

Confusion matrix를 그려본 결과 test set에 대해 위와 같은 결과를 도출하였다.

그러면 training 한 모델의 정확도를 평가하기 위한 지표를 계산해보자.

	TPR (Recall)	FPR	Precision	TNR	FNR	ACC	BCR	F1
Logstic Regression	0.6571429	0.01176471	0.9583333	0.9882353	0.3428571	0.8916667	0.8058609	0.779661

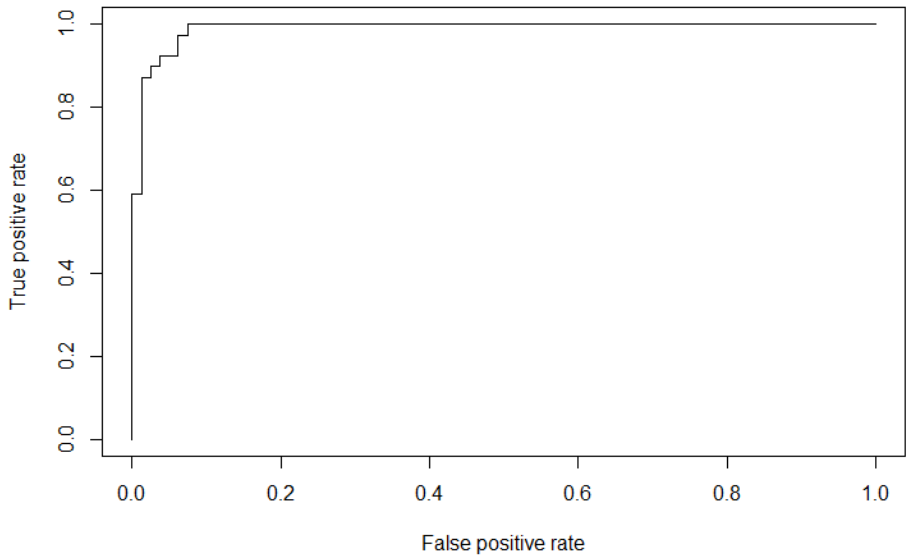
모델의 Performance 지표 값은 위와 같다.

TPR 은 실제 Positive 중 Positive로 분류한 비율, FPR 은 실제 negative인데 Positive로 잘못 분류한 비율, Precision은 예측이 positive라 한 것 중 실제 positive인 비율, TNR 은 실제 Negative인 것들 중 negative라 판정한 비율, FNR 은 실제 positive인데 negative라고 잘못 분류한 비율,

simple accuracy는 전체 값 중 맞춘 값의 비율이며, BCR과 F1은 ACCURACY 지표의 문제를 보완하기 위해 만든 새로운 ACCURACY 지표이다.

BCR은 TPR 과 TNR의 기하평균으로 계산하며, F1은 2*TPR*PRE/(TPR+PRE)로 계산한다.

<random seed 5회 변경 -> auroc 값 산출하기>



위 training한 모델의 ROC CURVE를 그려보면 위와 같은 결과가 나온다.

Seed를 5번 정도 변경해서 돌려도 비슷한 결과가 도출된다.

정확한 값 파악을 위해 AUROC 값을 측정하였다.

SEED에 따라 다르게 나온 결과들은 다음과 같다.

0.9912605, 0.9898702, 0.9737259, 0.9887566, 0.9785086 (1에 가까울수록 좋다)

TRAINING/TEST SET이 변경되어도 결과값들이 거의 동일함을 확인할 수 있다.

<Extra question>

<데이터 셋>

다양한 입력변수들과 이 입력변수에 따른 심장병 발병 유무(Y)가 기록된 데이터 셋이다.

입력변수는 AGE, SEX, CP, TRESTBPS, CHOL, FBS, RESTECG, THALACH, EXANG, OLDPEAK, SLOPE, CA,

THAL이 있으며 TARGET (Y값)이 0 or 1 이진수로 존재한다.

<training set, test set으로 분할 후 Logistic regression 모델 학습>

Training set과 test set을 7대 3으로 분류한 후 logistic regression 모델을 학습시켰다.

그 결과 다음과 같다.

```
Call:
glm(formula = target ~ ., family = "binomial", data = EX_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0854  -0.1211   0.0668   0.3312   2.2764

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.198801   4.812817  -0.457 0.647769
癤쭬ge        0.067544   0.037564   1.798 0.072162 .
sex1        -2.768089   0.913618  -3.030 0.002447 **
cp1          1.575517   0.850590   1.852 0.063988 .
cp2          2.276532   0.760436   2.994 0.002756 **
cp3          2.832883   1.026234   2.760 0.005772 **
trestbps    -0.051075   0.019342  -2.641 0.008274 **
chol        -0.002972   0.007211  -0.412 0.680217
fbs1         0.686081   0.970017   0.707 0.479388
restecg1     0.721930   0.576538   1.252 0.210504
restecg2    -0.999941   3.114234  -0.321 0.748144
thalach      0.049406   0.020258   2.439 0.014733 *
exang1      -1.550879   0.690820  -2.245 0.024769 *
oldpeak     -0.428393   0.359758  -1.191 0.233740
slope1      -1.027432   1.390843  -0.739 0.460082
slope2       0.330257   1.468668   0.225 0.822081
ca1         -3.105195   0.842275  -3.687 0.000227 ***
ca2         -5.912269   1.404327  -4.210 2.55e-05 ***
ca3         -2.227762   1.051229  -2.119 0.034074 *
ca4          1.669374   2.541778   0.657 0.511326
thal1        3.329039   2.961369   1.124 0.260947
thal2        1.943478   2.783626   0.698 0.485063
thal3        0.531851   2.791654   0.191 0.848906
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 278.048  on 201  degrees of freedom
Residual deviance:  97.945  on 179  degrees of freedom
AIC: 143.94

Number of Fisher Scoring iterations: 7
```

유의수준 0.05 수준에서 유의미한 변수를 선택해보자.

SEX(성별), CP, TRESTBPS, THALACH, EXANG1, CA 이렇게 6가지가 있다.

<CONFUSION MATRIX 그리기, PERFORMANCE 확인>

```
             EX_lr_predicted
EX_lr_target  0  1
           0 33  6
           1 10 38
```

CUT-OFF VALUE를 0.5로 잡았다.

CONFUSION MATRIX를 그린 결과 위와 같았다.

TRAIN한 모델이 어느 정도의 성능인지 PERFORMANCE를 확인해보자.

	TPR (Recall)	FPR	Precision	TNR	FNR	ACC	BCR	F1
Logstic Regression	0.7916667	0.1538462	0.8636364	0.8461538	0.2083333	0.816092	0.818457	0.826087

BCR과 F1값이 SIMPLE ACCURACY 값과 큰 차이가 없다.

그리고 그 값들이 0.81 이상으로 높다. 그래서 TRAINING한 모델이 좋다고 말할 수 있다.

그런데, 위의 데이터는 의료, 생명 관련 데이터이다. 즉, POSITIVE(정상이 아닌 것)를 판별하는 CUT-OFF VALUE를 엄격하게(낮게) 잡을 필요가 있다.

그래서 CUT-OFF VALUE를 0.2로 잡고 다시 CONFUSION MATRIX와 PERFORMANCE 지표를 도출하였다.

```
             EX_lr_predicted
EX_lr_target  0  1
           0 30  9
           1  5 43
```

그 결과 CONFUSION MATRIX는 위와 같았다.

실제 값이 POSITIVE인 것을 NEGATIVE로 잘못 예측하는 수가 줄어들었다.

	TPR (Recall)	FPR	Precision	TNR	FNR	ACC	BCR	F1
Logstic Regression	0.8958333	0.2307692	0.8269231	0.7692308	0.1041667	0.8390805	0.830122	0.86

그리고 PERFORMANCE 값도 조금 더 향상한 것을 확인할 수 있다.

그래서 CUT-OFF VALUE를 낮게 잡은 이 결과값이 위보다 더 합당할 것이다.