Multivariate Data Analysis
Assignment #4

Logistic Regression: Graduate Admissions

과목명	다변량분석
담당교수	강필성 교수님
제출일	2019-05-16
이름	박지원
학과명	산업경영공학부
학번	2014170856

목차

1. GRADUATE ADMISSIONS DATASET의 LOGISTIC REGRESSION MODEL · · · · · · · · 3
1.1. [Q1] 모델 구축에 필요하지 않은 변수 · · · 3
1.2. [Q2] 입력 변수들의 단변량 통계량과 Boxplot · · · 3
1.3. [Q3] 이상치의 조건 정의와 제거 · · · 4
1.4. [Q4] Scatterplot과 Correlation plot · · · 4
1.5. [Q5] 학습 데이터로 Logistic Regression 모델 구축 · · · 5
1.5.1 종속변수 변환과 입력변수의 normalization · · · 5
1.5.2. 학습 데이터로 모델 구축 · · · 5
1.6. [Q6] Test 데이터셋에 대한 예측과 Confusion Matrix · · · 6
1.6.1. Confusion Matrix · · · 6
1.6.2. Performance Measures · · · 6
1.7. [Q7] AUROC 산출 · · · 7
1.7.1. Seed = 12345 · · · 7
1.7.2. Seed = 27153 · · · 7
1.7.3. Seed = 12553 · · · 8
1.7.4. Seed = 16373 · · · 8
1.7.5. Seed = 92735 · · · 8
1.7.6. AUROC 비교 · · · 9
2. [Q8] 기타 DATASET의 LOGISTIC REGRESSION MODEL · · · 10
2.1. 모델 구축에 필요하지 않은 변수 · · · 10
2.2. 입력 변수들의 단변량 통계량과 Boxplot · · · 10
2.3. 이상치의 조건 정의와 제거 · · · 11
2.4. Scatterplot과 Correlation plot · · · 11
2.5. 학습 데이터로 Logistic Regression 모델 구축 · · · 12
2.5.1 입력변수의 normalization · · · 12
2.5.2. 학습 데이터로 모델 구축 · · · 13
2.6. Test 데이터셋에 대한 예측과 Confusion Matrix · · · 13
2.6.1. Confusion Matrix · · · 13
2.6.2. Performance Measures · · · 14
2.7. AUROC 산출 · · · 14
2.7.1. Seed = 12345 · · · 14
2.7.2. Seed = 39157 · · · 15
2.7.3. Seed = 20375 · · · 15
2.7.4. Seed = 39472 · · · 15
2.7.5. Seed = 71526 · · · 16
2.7.6. AUROC 비교 · · · 16

1. Graduate Admissions Dataset의 Logistic Regression Model

Dataset: Graduate Admissions
이 데이터셋은 미국 대학원의 지원자들에 대한 여러 가지 점수(GRE, TOELF 등)와 대학의 등급에 따라 각 지원자들이 합격할 확률(Chance of Admit)을 기록한 데이터이다. X 변수들을 이용해 합격 확률을 예측하고자 한다.

1.1. [Q1] 모델 구축에 필요하지 않은 변수

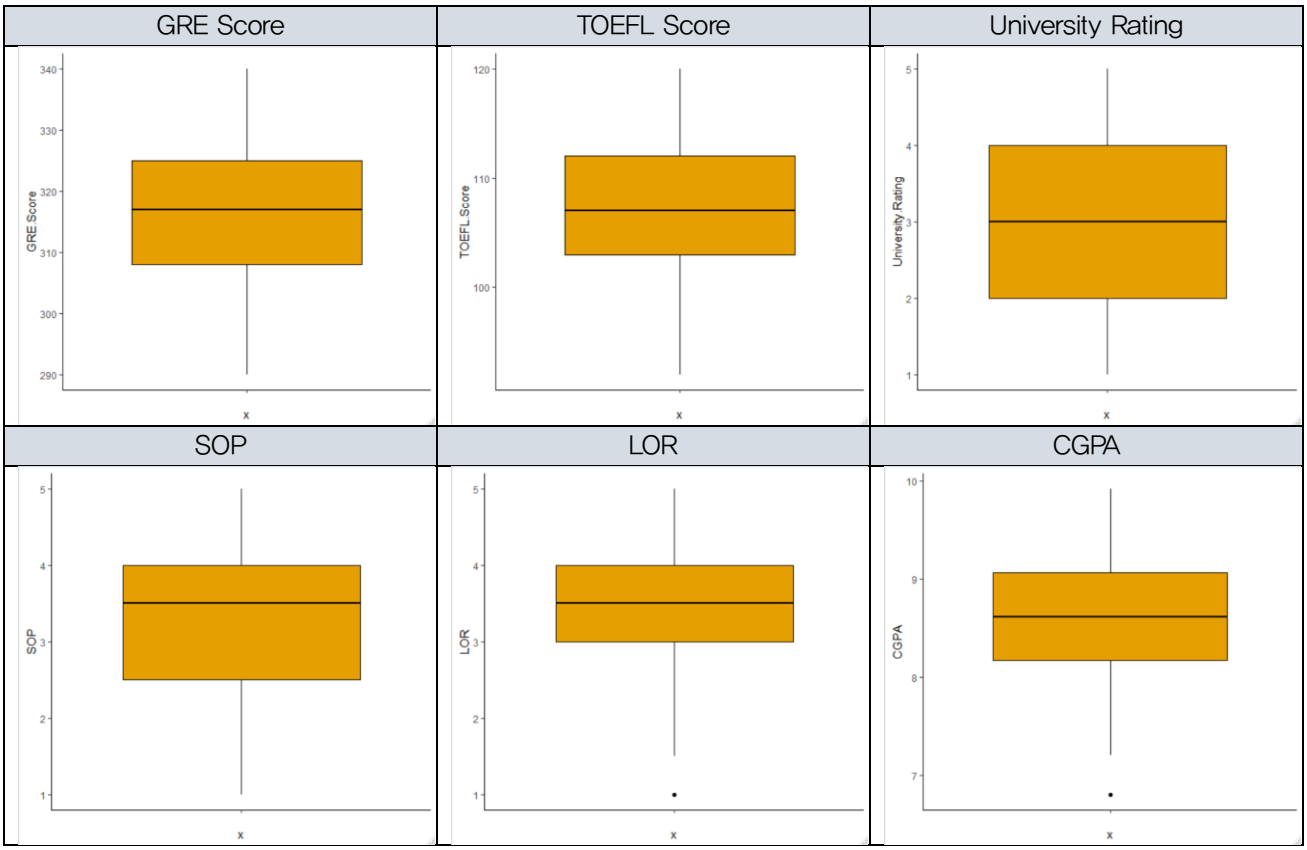
Serial Number는 Logistic Regression 모형 구축을 위해 필요하지 않다. 서로 다른 데이터를 식별하는 코드일 뿐이기 때문이다.

1.2. [Q2] 입력 변수들의 단변량 통계량과 Boxplot

개별 입력 변수들에 대한 Mean, Standard Deviation, Skewness, Kurtosis와 Box plot은 아래와 같다. 1.1에서 필요하지 않다고 판단한 serial number 변수는 제거했다.

Variable	Mean	Standard Deviation	Skewness	Kurtosis
GRE.Score	316.8075	11.47365	−0.06242	−0.71818
TOEFL.Score	107.41	6.069514	0.056788	−0.59858
University.Rating	3.0875	1.143728	0.169978	−0.81231
SOP	3.4	1.006869	−0.2737	−0.69373
LOR	3.4525	0.898478	−0.10619	−0.68083
CGPA	8.598925	0.596317	−0.0655	−0.48037

Table 1 입력 변수들의 단변량 통계량



전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 다음과 같다.

(1) GRE Score: Box plot과 skewness로부터 대칭성을 확인할 수 있었다. 단, kurtosis가 −0.7181로 정규분포가 갖는 kurtosis (3)보다 작은 값이었다. 정규분포보다 약간 납작한 모양일 것으로 생각된다.
(2) TOEFL Score: Box plot이 대칭이고, skewness가 매우 작으므로 대칭이라고 가정했다. 마찬가지로 kurtosis가 음수로, 정규분포보다 아래로 눌린 모양일 것이다.
(3) CGPA: Box plot과 skewness를 살펴보았을 때, 데이터가 대칭이라고 생각했다. 위의 두 변수와 마찬가지로 negative kurtosis를 가지므로 정규분포보다 뭉툭한 모양일 것이다.

1.3. [Q3] 이상치의 조건 정의와 제거

위의 Boxplot들을 봤을 때, GRE score, TOEFL score, University Rating, SOP, 그리고 Research는 이상치가 발견되지 않았다. LOR과 CGPA는 [(Q1−1.5*IQR), (Q3+1.5*IQR)]를 벗어난 범위에서 이상치가 발견되었다. 이에 해당하는 객체들을 데이터셋에서 제거했다.

1.4. [Q4] Scatterplot과 Correlation plot

수치형 입력 변수들의 조합에 대해 Scatterplot과 Correlation plot을 그려보았다.

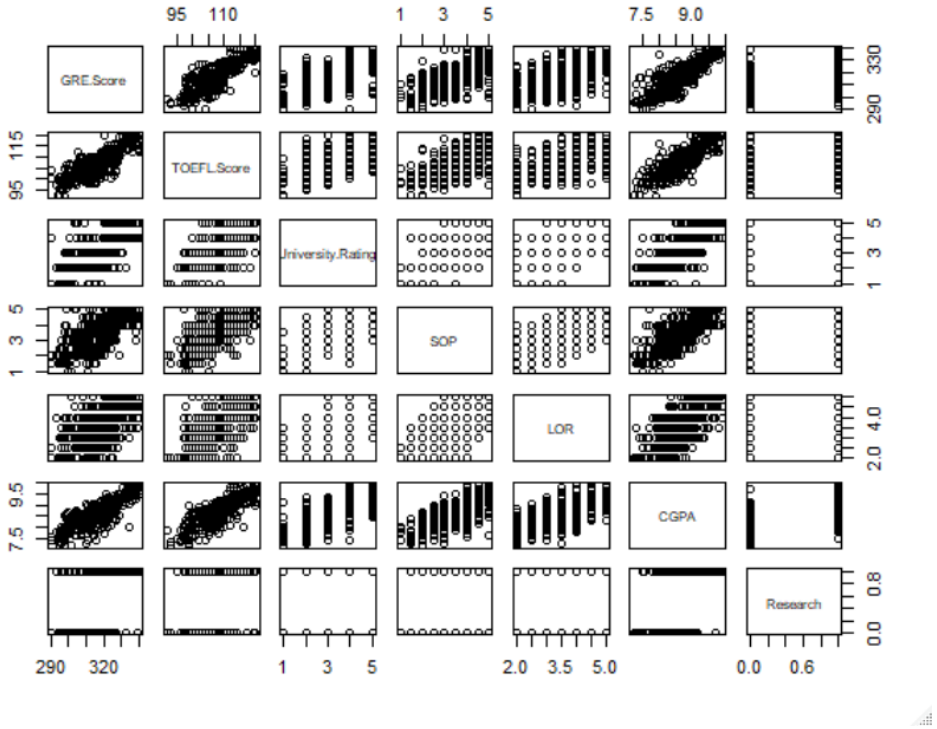


Figure 1 입력 변수들의 Scatterplot

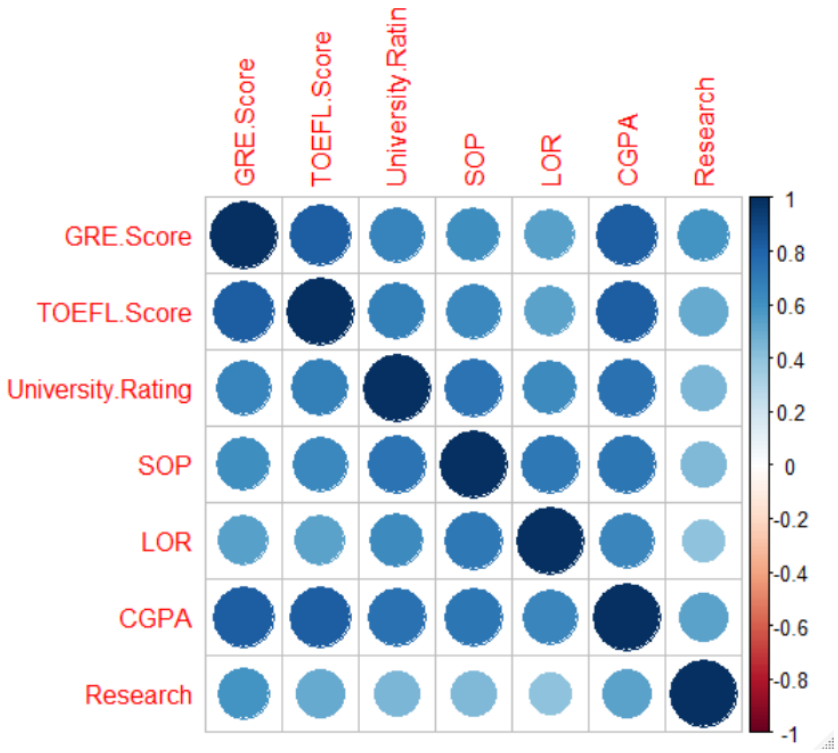


Figure 2 입력 변수들의 Correlation plot

Figure 2의 correlation plot으로부터 GRE Score와 TOEFL Score, 그리고 GRE Score와 CGPA가 가장 큰 상관관계를 가지는 것을 알 수 있다. Scatter plot에서도 양의 상관관계가 뚜렷하게 드러났다. 시험 성적이 좋은 학생이 다른 시험이나 학점에서도 좋은 점수를 받기 때문으로 추측된다. CGPA는 GRE Score 뿐만 아니라 TOEFL Score, University Rating, SOP 등 다른 변수들과도 강한 양의 상관관계를 갖는 것으로 보인다.

1.5. [Q5] 학습 데이터로 Logistic Regression 모델 구축

1.5.1 종속변수 변환과 입력변수의 normalization

종속변수인 Chance of Admit을 0.8을 초과하는지 여부로 1 (positive class)과 0 (negative class)의 값을 갖는 binary target variable로 변환했다. 다음으로, 입력변수들에 대해 normalization을 수행했다. 이론적으로는 logistic regression에서 normalization이 필수적인 것은 아니지만, 입력변수끼리 값의 범위가 크게 차이 날 경우, R에서 계산하며 rounding error가 발생할 수 있기 때문이다.

1.5.2. 학습 데이터로 모델 구축

Seed를 12345로 설정하여 전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 분할한 후 모든 변수를 사용하여 Logistic Regression 모델을 학습했다. 결과는 다음과 같다.

```
Call:
glm(formula = admission_target ~ ., family = binomial, data = admission_trn)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-2.27710  -0.11104  -0.01289   0.03984   2.53912
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.8263     0.6485  -5.901 3.62e-09 ***
GRE.Score         -0.9415     0.7661  -1.229   0.2191
TOEFL.Score        1.1058     0.6300   1.755   0.0792 .
University.Rating  1.2096     0.4881   2.478   0.0132 *
SOP               -1.0566     0.6090  -1.735   0.0828 .
LOR                0.7699     0.4300   1.790   0.0734 .
CGPA               5.1504     1.1692   4.405 1.06e-05 ***
Research           0.9350     0.3883   2.408   0.0160 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 332.695  on 273  degrees of freedom
Residual deviance:  80.393  on 266  degrees of freedom
AIC: 96.393

Number of Fisher Scoring iterations: 8
```

유의수준 0.1에서 Chance of Admit에 유의미하게 영향을 주는 변수들을 파악하기 위해 p-value를 살펴보았다. Coefficient의 값이 0이라는 귀무가설을 기각하기 위해서는 p-value가 0.1보다 낮아야 한다. P-value가 0,1보다 작은, 즉 유의미한 변수들은 TOEFL Score, University.Rating, SOP, LOR, CGPA, Research였다.

1.6. [Q6] Test 데이터셋에 대한 예측과 Confusion Matrix

1.6.1. Confusion Matrix

Test 데이터셋에 대하여 예측을 수행하고 Confusion Matrix를 생성했다. Cutoff는 0.8로 설정했다.

		Predicted	
		0	1
Target	0	78	4
	1	7	28

모델이 TEST 데이터셋에 대해 합격을 제대로 예측한 경우가 28번, 불합격을 제대로 예측한 경우가 78번, 합격인데 불합격이라고 잘못 예측한 경우가 7번, 불합격인데 합격이라고 잘못 예측한 경우가 4번이었다.

1.6.2. Performance Measures

위의 Confusion matrix를 바탕으로 True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, Simple Accuracy, Balanced Correction Rate, F1-Measure를 구했다. 결과는 아래와 같다.

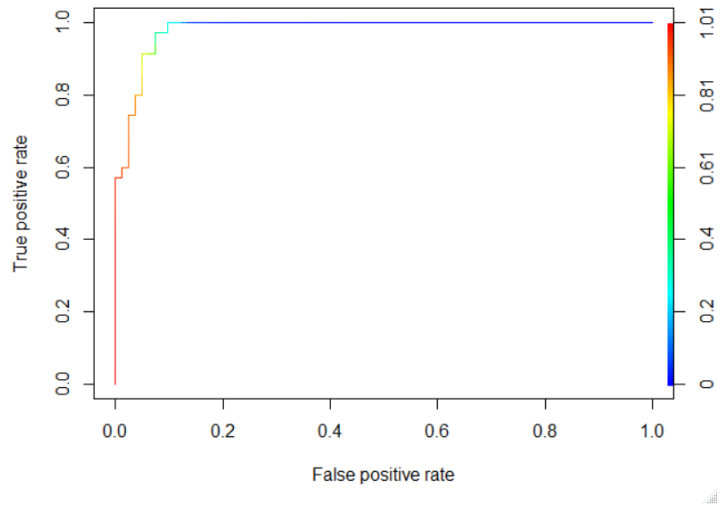
TPR	TNR	FPR	FNR	ACC	BCR	PRE	F1
0.8	0.9512195	0.2	0.04878049	0.9059829	0.8723392	0.875	0.8358209

단순 accuracy는 0.9059로 높았다. TPR은 이보다는 낮은 0.8이었다. 이는 불합격 데이터가 합격 데이터에 비해 많았기 때문으로 예상된다. TPR이 0.8이라는 것은 실제 합격자 중 80%가 제대로 identified 되었음을 의미한다. Precision은 0.875로 모델에 의해 합격이라고 예측된 사람들 중 87.5%가 실제 합격자였음을 의미한다. 0.9512의 TNR은 실제 불합격자 중 95.12%가 제대로 identified 되었음을 나타낸다. 단순 accuracy만으로는 모델의 performance를 정확히 평가할 수 없다. 데이터의 불균형이 심한 경우 한 클래스를 제대로 예측하지 못하더라도 다른 클래스를 제대로 예측한 횟수가 많아서 높게 계산될 수 있기 때문이다. 따라서 이러한 단점을 보완한 BCR과 F1 measure도 살펴보아야 한다. BCR은 TNR과 TPR을 곱한 값에 루트를 씌운 것이다. 따라서 class imbalance가 있을 때 TNR이나 TPR이 작게 나온다면 전체 값이 작아지게 되어있다. F1은 Precision과 Recall의 가중평균으로, FP와 FN을 함께 고려한다. BCR은 0.8723392이었고 F1은 0.8358209이었다. 모두 높은 값을 가지므로 이 모델은 좋은 performance를 낸다고 할 수 있다.

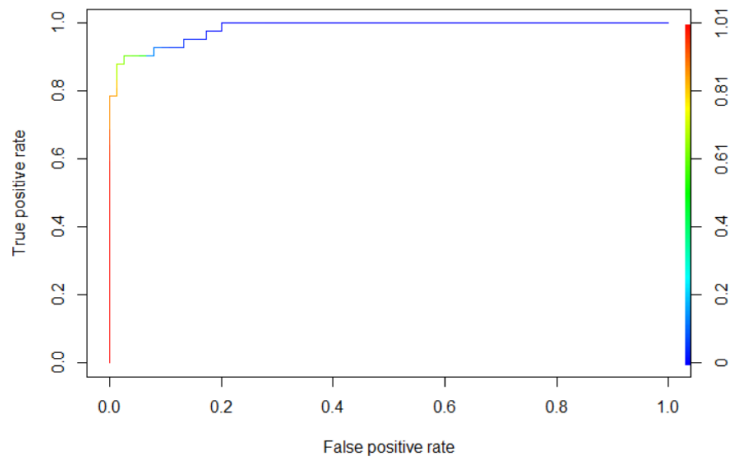
1.7. [Q7] AUROC 산출

Random seed를 변경해가며 학습 테스트를 5회 반복했다. 각 테스트에 대해 ROCR 라이브러리를 사용해ROC를 그리고 직접 AUROC를 산출했다. 결과는 아래와 같다.

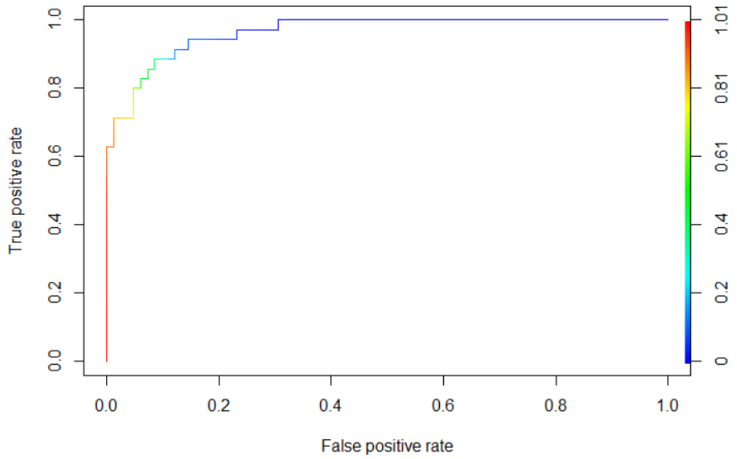
1.7.1. Seed = 12345



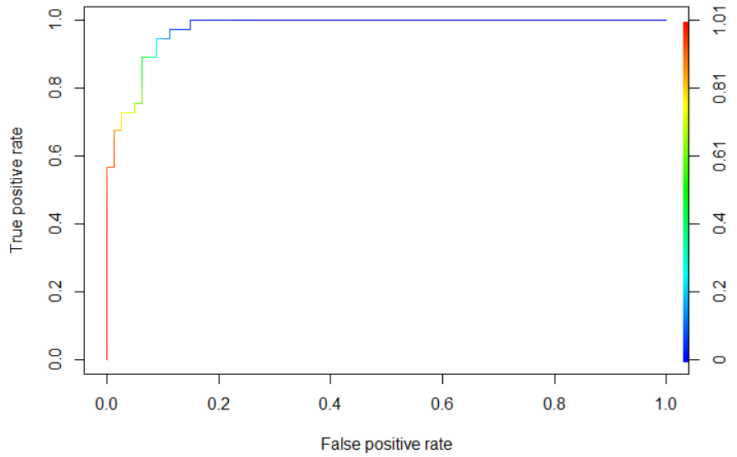
1.7.2. Seed = 27153



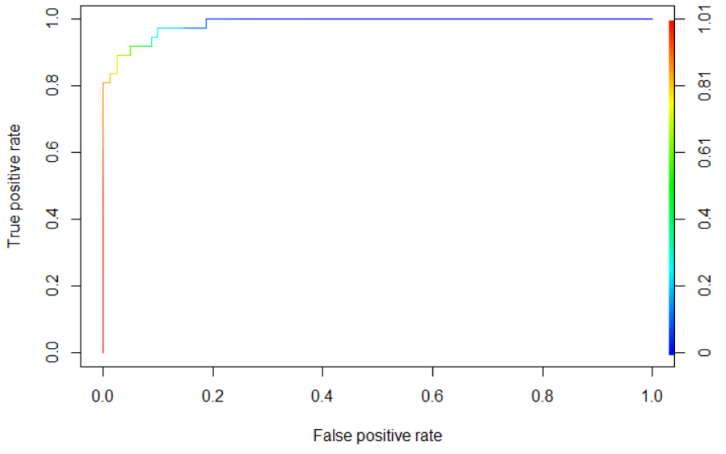
1.7.3. Seed = 12553



1.7.4. Seed = 16373



1.7.5. Seed = 92735



1.7.6. AUROC 비교

Seed	12345	27153	12553	16373	92735
AUROC	0.9815331	0.984127	0.9655052	0.9756757	0.9868243
Mean	0.9787331				

Seed가 달라질 때마다 training set과 testing set이 달라져 매번 다른 ROC가 그려지고, 다른 AUROC가 계산되었다. 5개의 AUROC 평균은 0.9787331이었다. False Positive Rate가 0이고, True Positive Rate가 1인 커브가 가장 이상적이므로, 이상적인 AUROC값은 1이다. 랜덤하게 예측했을 때는 커브가 대각선으로 그려지므로 AUROC 값은 0.5가 된다. 평균 AUROC가 0.9787331이라는 것은 1에 매우 가까운 수치이므로 이 모델은 좋은 classifier라고 할 수 있다.

ROC 커브로부터 최적의 cut-off 값도 찾아낼 수 있다. 왼쪽 위 모서리에 가까울수록 FPR은 0에, TPR은 1에 가깝다. 왼쪽 위 모서리에 가장 가까운 커브의 색깔을 살펴보면 대부분 노란색과 초록색이 섞여있다. 커브 오른쪽의 index를 보면 노란색과 초록색은 0.6~0.7 정도를 나타내므로, cut-off를 0.6~0.7로 설정하는 것이 타당하다.

2. [Q8] 기타 Dataset의 Logistic Regression Model

Dataset: Pima Indians Diabetes Database (https://www.kaggle.com/kumargh/pimaindiansdiabetescsv)
이 데이터셋은 피마 여성 인디언의 나이, 임신 횟수, 혈압, BMI 지수 등 건강 정보와 당뇨병 발생 여부를 기록한 데이터이다. 이 데이터셋으로 Logistic regression 모델을 구축해 당뇨병 발생 여부를 예측하고자 한다.

Variables	설명
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2–Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Outcome	Class variable (1:tested positive for diabetes, 0: tested negative for diabetes)

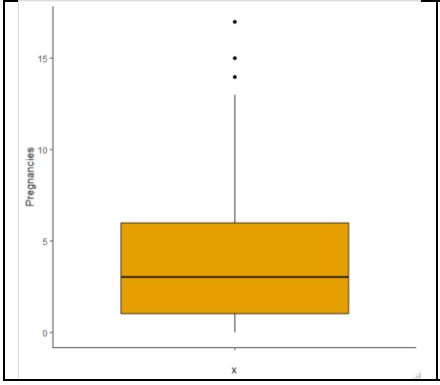
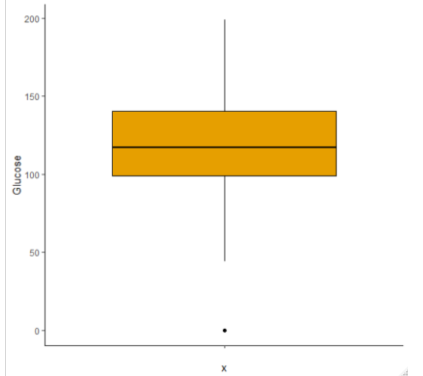
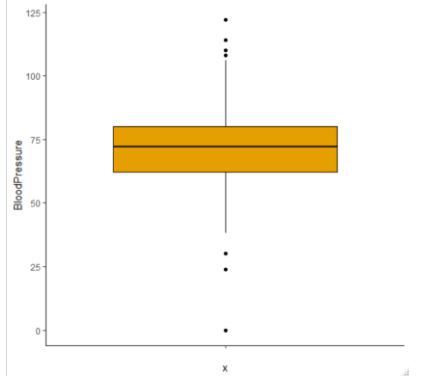
2.1. 모델 구축에 필요하지 않은 변수

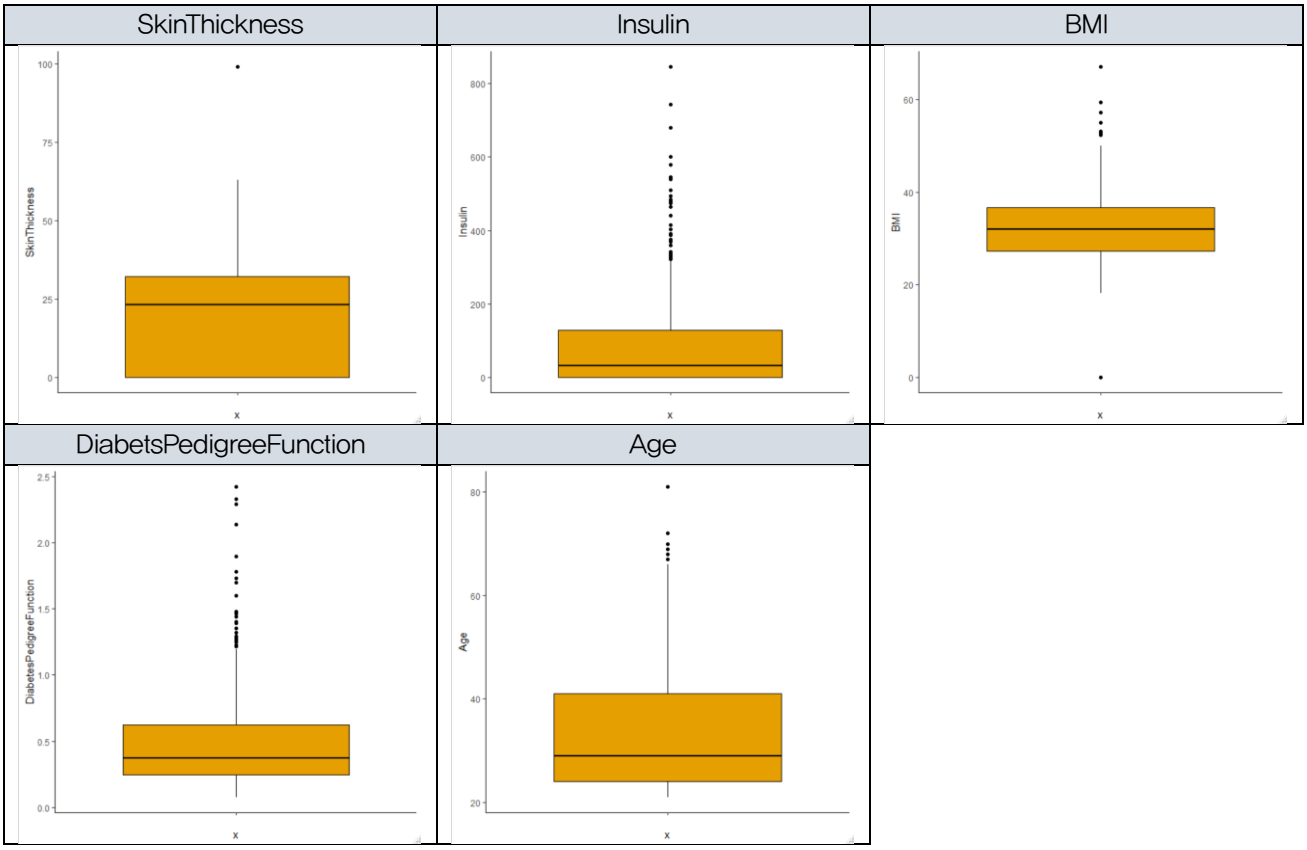
유의미하지 않은 변수가 없다고 판단해 모델 구축에 모두 사용하기로 하였다.

2.2. 입력 변수들의 단변량 통계량과 Boxplot

개별 입력 변수들에 대한 Mean, Standard Deviation, Skewness, Kurtosis와 Box plot은 아래와 같다.

Variable	Mean	Standard Deviation	Skewness	Kurtosis
Pregnancies	3.845052	3.369578	0.898155	0.142184
Glucose	120.8945	31.97262	0.173075	0.619369
BloodPressure	69.10547	19.35581	−1.83641	5.11751
SkinThickness	20.53646	15.95222	0.108946	−0.53094
Insulin	79.79948	115.244	2.263383	7.133135
BMI	31.99258	7.88416	−0.42731	3.244963
DiabetesPedigreeFunction	0.471876	0.331329	1.912418	5.528539
Age	33.24089	11.76023	1.125188	0.621727

Pregnancies	Glucose	BloodPressure
		



전체 변수 중에서 정규분포를 따른다고 할 수 있는 변수들은 다음과 같다.

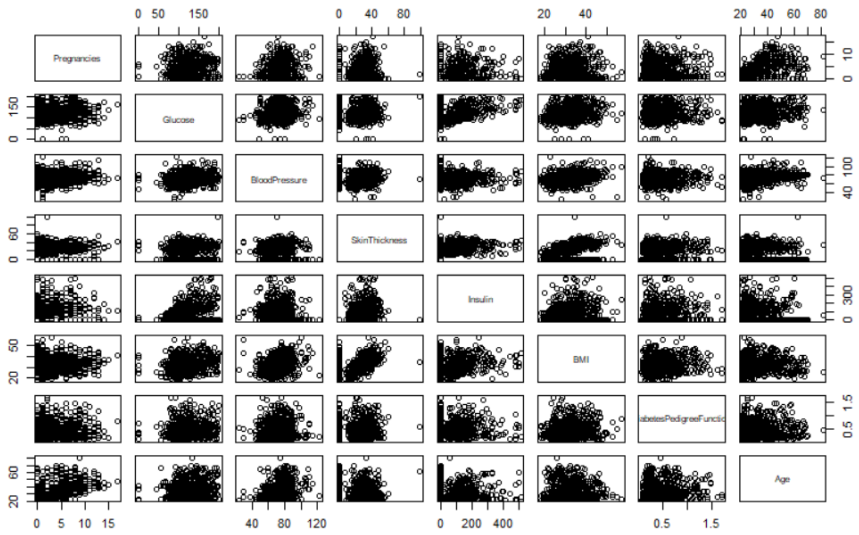
(1) Glucose: Box plot과 skewness로부터 대칭성은 확인했지만 kurtosis가 0.619369로 정규분포가 갖는 kurtosis보다 작은 값이었다.
(2) BloodPressure: Box plot이 대칭이고, skewness가 매우 작아 대칭일 것으로 가정했다. kurtosis가 5.11751로, 정규분포보다 살짝 뾰족한 모양일 것이다.
(3) BMI: Box plot과 skewness를 보고 데이터가 대칭이라고 생각했다. Kurtosis 역시 3.244963으로, 정규분포의 kurtosis인 3과 매우 비슷했다.

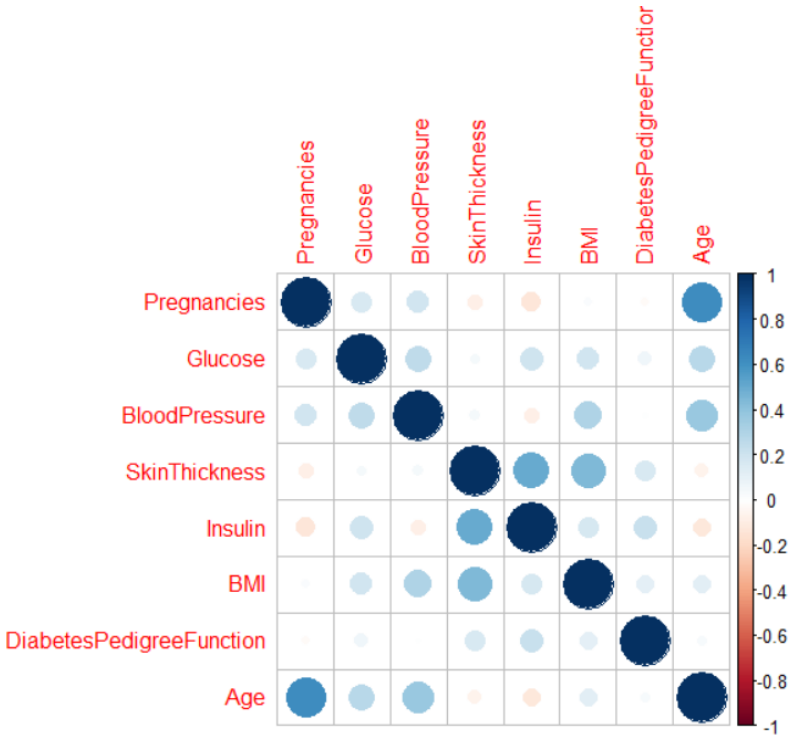
2.3. 이상치의 조건 정의와 제거

위의 Boxplot으로부터 대다수의 변수에 이상치가 존재함을 알 수 있었다. 특히 DiabetesPedigreeFunction과 Insulin은 이상치가 매우 많이 관측되었다. [(Q1−1.5*IQR), (Q3+1.5*IQR)]를 벗어난 범위의 이상치를 모두 제거할 경우 너무 많은 수의 데이터가 제거될 것으로 예상되어 [(Q1−3*IQR), (Q3+3*IQR)]을 벗어난 데이터만을 이상치로 정의하고 제거했다.

2.4. Scatterplot과 Correlation plot

수치형 변수들의 조합에 대해 Scatterplot과 Correlation plot을 그려보았다. 결과는 아래와 같다.





변수들 중에서 Age와 Pregnancies가 가장 큰 양의 상관관계를 갖는 것으로 확인되었다. 다음으로 Insulin과 SkinThickness의 상관관계가 컸고, BMI와 SkinThickness가 다음으로 큰 상관관계를 가졌다.

2.5. 학습 데이터로 Logistic Regression 모델 구축

2.5.1 입력변수의 normalization

입력변수들에 대해 normalization을 수행했다. 이론적으로는 logistic regression에서 normalization이 필수적인 것은 아니지만, 입력변수끼리 값의 범위가 크게 차이 날 경우, R에서 계산하며 rounding error가 발생할 수 있기 때문이다.

2.5.2. 학습 데이터로 모델 구축

전체 데이터셋을 70%의 학습 데이터와 30%의 테스트 데이터로 무작위로 분할한 후 모든 변수를 사용하여 Logistic Regression 모델을 학습했다. 결과는 다음과 같다.

```
Call:
glm(formula = diabetes_target ~ ., family = binomial, data = diabetes_trn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6856  -0.6906  -0.3724   0.6788   2.6720

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.9911     0.1268  -7.813 5.59e-15 ***
Pregnancies                0.4253     0.1373   3.096  0.00196 **
Glucose                    1.1705     0.1521   7.693 1.44e-14 ***
BloodPressure             -0.1331     0.1269  -1.048  0.29447
SkinThickness              0.1354     0.1437   0.942  0.34609
Insulin                   -0.2202     0.1449  -1.519  0.12870
BMI                        0.7996     0.1493   5.356 8.50e-08 ***
DiabetesPedigreeFunction   0.2983     0.1183   2.522  0.01167 *
Age                        0.2140     0.1409   1.518  0.12896
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.70  on 499  degrees of freedom
Residual deviance: 445.19  on 491  degrees of freedom
AIC: 463.19

Number of Fisher Scoring iterations: 5
```

유의수준 0.1에서 당뇨병 발병에 유의미하게 영향을 주는 변수들을 파악하기 위해 p-value를 살펴보았다. Coefficient의 값이 0이라는 귀무가설을 기각하기 위해서는 p-value가 0.1보다 작아야 한다. 이에 해당하는 변수들은 Pregnancies, Glucose, BMI, DiabetesPedigreeFunction이었다.

2.6. Test 데이터셋에 대한 예측과 Confusion Matrix

2.6.1. Confusion Matrix

Test 데이터셋에 대하여 예측을 수행하고 Confusion Matrix를 생성했다. Cutoff는 0.5로 설정했다.

		Predicted	
		0	1
Target	0	121	22
	1	39	33

모델이 TEST 데이터셋에 대해 당뇨병 발병을 제대로 예측한 경우가 33번, 발병하지 않음을 제대로 예측한 경우가 121번, 발병했는데 하지 않았다고 잘못 예측한 경우가 39번, 발병하지 않았는데 발병했다고 잘못 예측한 경우가 22번이었다.

2.6.2. Performance Measures

위의 Confusion matrix를 바탕으로 True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, Simple Accuracy, Balanced Correction Rate, F1-Measure를 구했다. 결과는 아래와 같다.

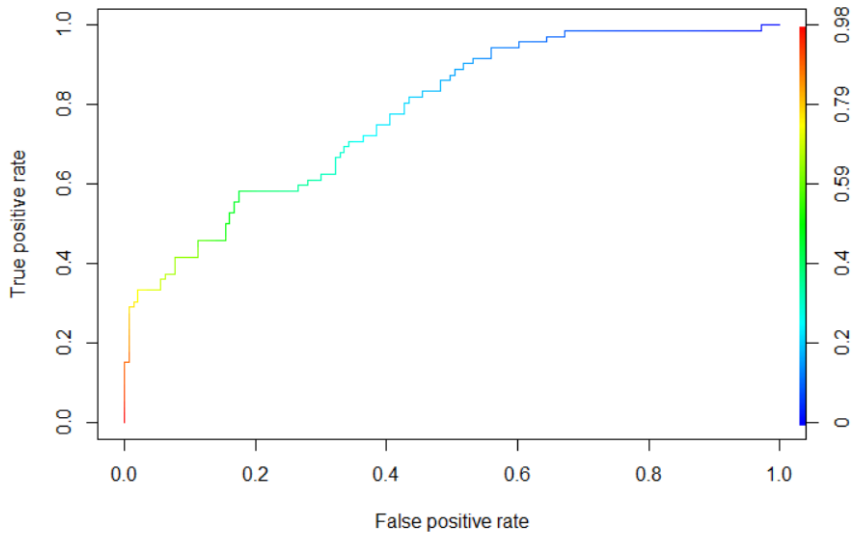
TPR	TNR	FPR	FNR	ACC	BCR	PRE	F1
0.4583333	0.8461538	0.5416667	0.1538462	0.7162791	0.6227524	0.6	0.519685

단순 accuracy는 0.7163으로 높은 편이었다. 하지만 TPR은 이보다 낮은 0.4583이었다. 이는 당뇨병 발병을 제대로 예측하지 못했음에도 불구하고 당뇨병 미발병 클래스가 발병 클래스보다 훨씬 많았기 때문으로 예상된다. Confusion matrix에서도 대다수의 당뇨병 미발병 클래스는 제대로 분류된 것에 비해 발병 클래스는 제대로 분류된 것보다 제대로 분류되지 않은 경우가 많음을 알 수 있다. 이는 TPR과 TNR을 통해 수치로 확인할 수 있다. TPR이 0.4583이라는 것은 실제 발병자 중 45.83%만이 제대로 분류되었음을 의미하고 TNR이 0.8461이라는 것은 실제 미발병자 중 84.61%가 제대로 분류되었음을 나타내기 때문이다. 이렇게 발병 클래스에 대해 예측을 제대로 못했기 때문에 BCR과 F1 역시 낮게 산출되었음을 알 수 있다.

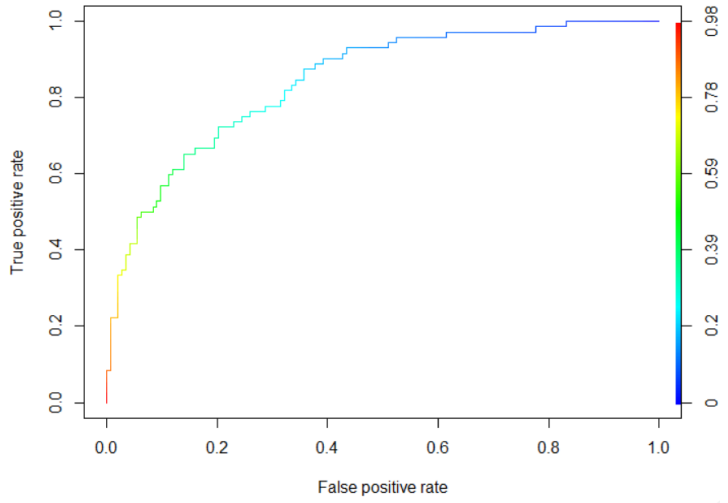
2.7. AUROC 산출

Random seed를 변경해가며 학습 테스트를 5회 반복했다. 각 테스트에 대해 ROCR 라이브러리를 사용해ROC를 그리고 직접 AUROC를 산출했다. 결과는 아래와 같다.

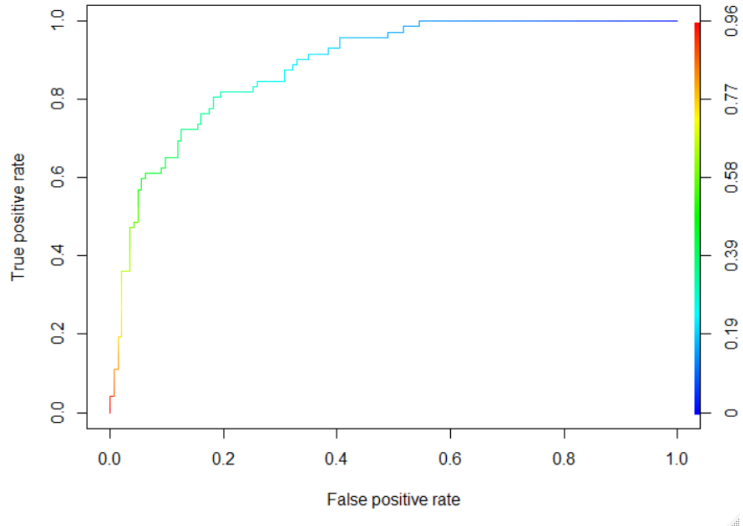
2.7.1. Seed = 12345



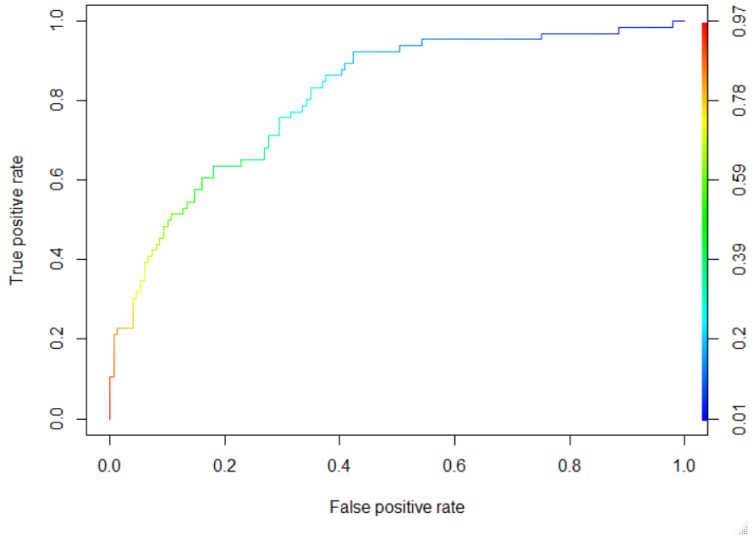
2.7.2. Seed = 39157



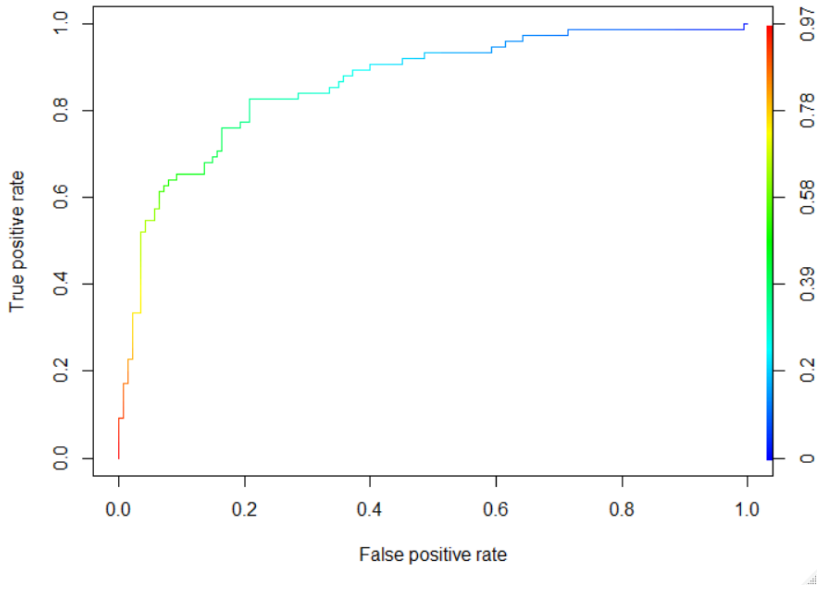
2.7.3. Seed = 20375



2.7.4. Seed = 39472



2.7.5. Seed = 71526



2.7.6. AUROC 비교

Seed	12345	39157	20375	39472	71526
AUROC	0.7797203	0.8438228	0.8869464	0.8105552	0.8632381
Mean	0.8368566				

Seed가 달라질 때마다 training set과 testing set이 달라져 매번 다른 ROC가 그려지고, 다른 AUROC가 계산되었다. 5개의 AUROC 평균은 0.8368566이었다. 이상적 수치인 1에 가까우므로 random classifier보다는 좋은 classifier임을 알 수 있다.

ROC 커브로부터 최적의 cut-off 값도 찾아낼 수 있다. 왼쪽 위 모서리에 가까울수록 FPR은 0에, TPR은 1에 가깝다. 왼쪽 위 모서리에 가장 가까운 커브의 색깔을 살펴보면 대부분 초록색이다. 커브 오른쪽의 index에 따르면 초록색은 0.4~0.6 정도를 나타내므로, cut-off를 0.4~0.6로 설정하는 것이 타당하다.