다변량 분석 Assignment 4

2014170824 황태민

[Q1]

Kaggle에서 소개하는 변수들의 설명은 다음과 같다.

'The dataset contains several parameters which are considered important during the application for Masters Programs. The parameters included are : 1. GRE Scores (out of 340) 2. TOEFL Scores (out of 120) 3. University Rating (out of 5) 4. Statement of Purpose and Letter of Recommendation Strength (out of 5) 5. Undergraduate GPA (out of 10) 6. Research Experience (either 0 or 1) 7. Chance of Admit (ranging from 0 to 1)'

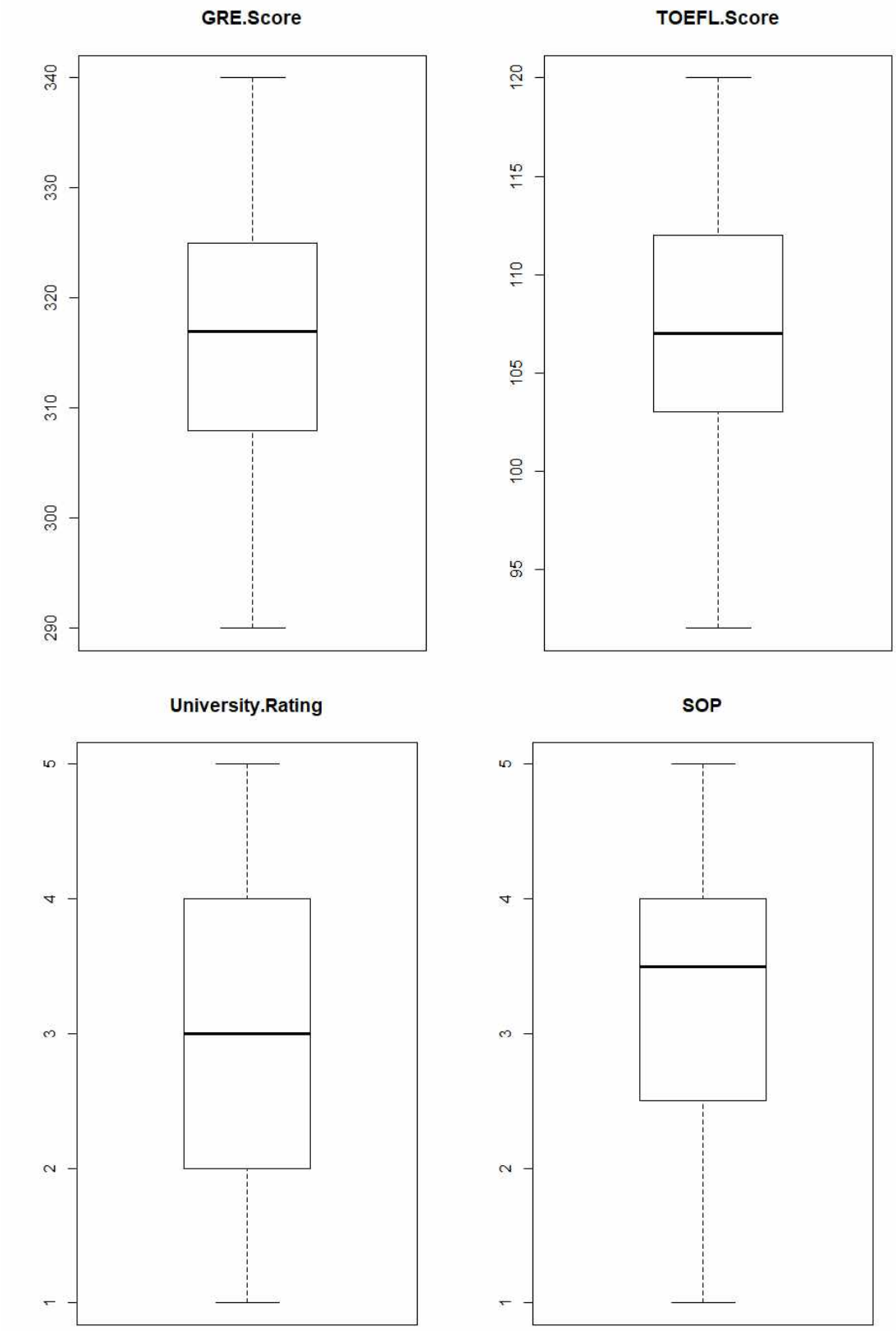
Logistic Regression 모형 구축에 Serial No. 변수는 필요하지 않다. 단지 순서만을 나타내는 변수이기 때문에 입력변수로서 사용되어선 안된다. 이에 따라 Chance of Admit 변수를 종속 변수로 놓고 나머지 변수들을 입력 변수로 놓고 모델을 세웠다.

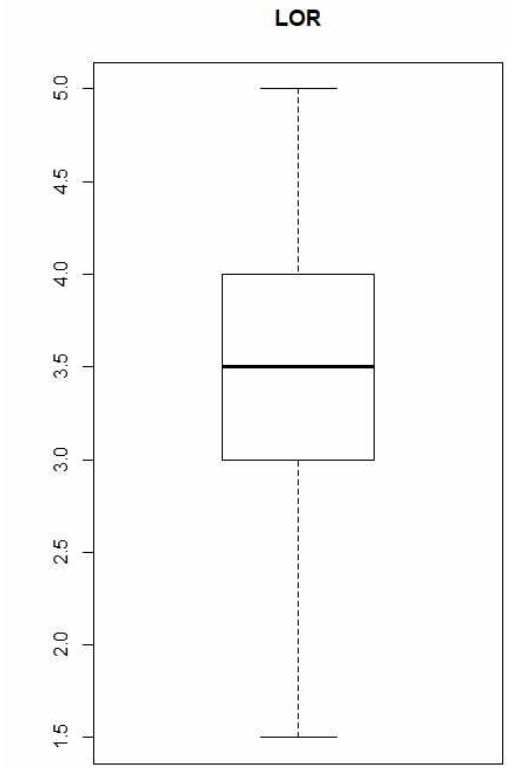
[Q2]

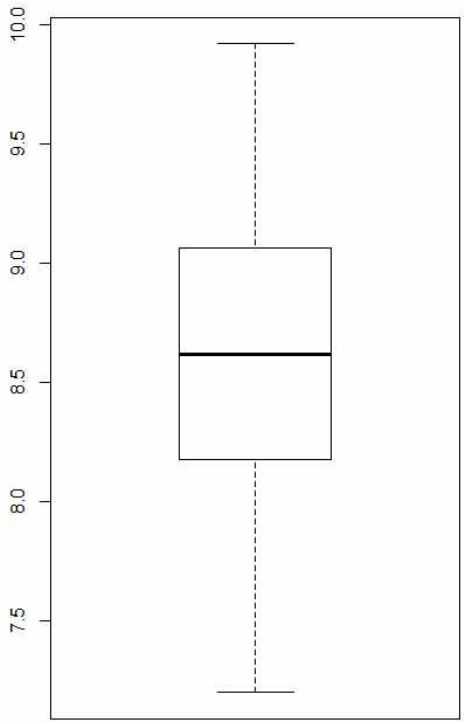
각 입력 변수의 Mean, Standard Deviation, Skewness, Kurtosis 값은 다음과 같다.

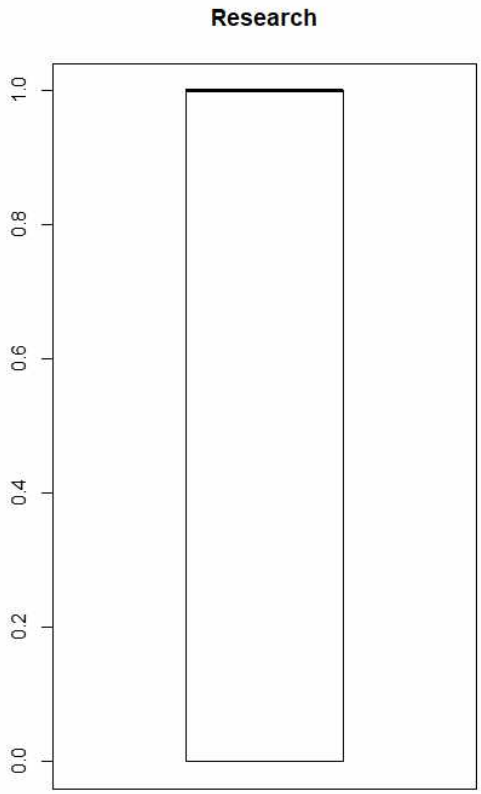
	Mean	Standard Deviation	Skewness	Kurtosis
GRE.Score	316.807500	11.4736461	-0.06265736	2.293273
TOEFL.Score	107.410000	6.0695138	0.05700113	2.413468
University.Rating	3.087500	1.1437281	0.17061738	2.198669
SOP	3.400000	1.0068686	-0.27472598	2.317843
LOR	3.452500	0.8984775	-0.10658984	2.330805
CGPA	8.598925	0.5963171	-0.06574282	2.532273
Research	0.547500	0.4983620	-0.19086322	1.036429

각 입력 변수의 box plot은 다음과 같다.









각 변수들이 정규분포를 따르기 위해서는 Skewness의 값이 0에 가깝고, Kurtosis의 값이 3에 가까워야 한다. 7개의 변수 모두 Skewness의 값이 0에서 크게 벗어나지 않았고,

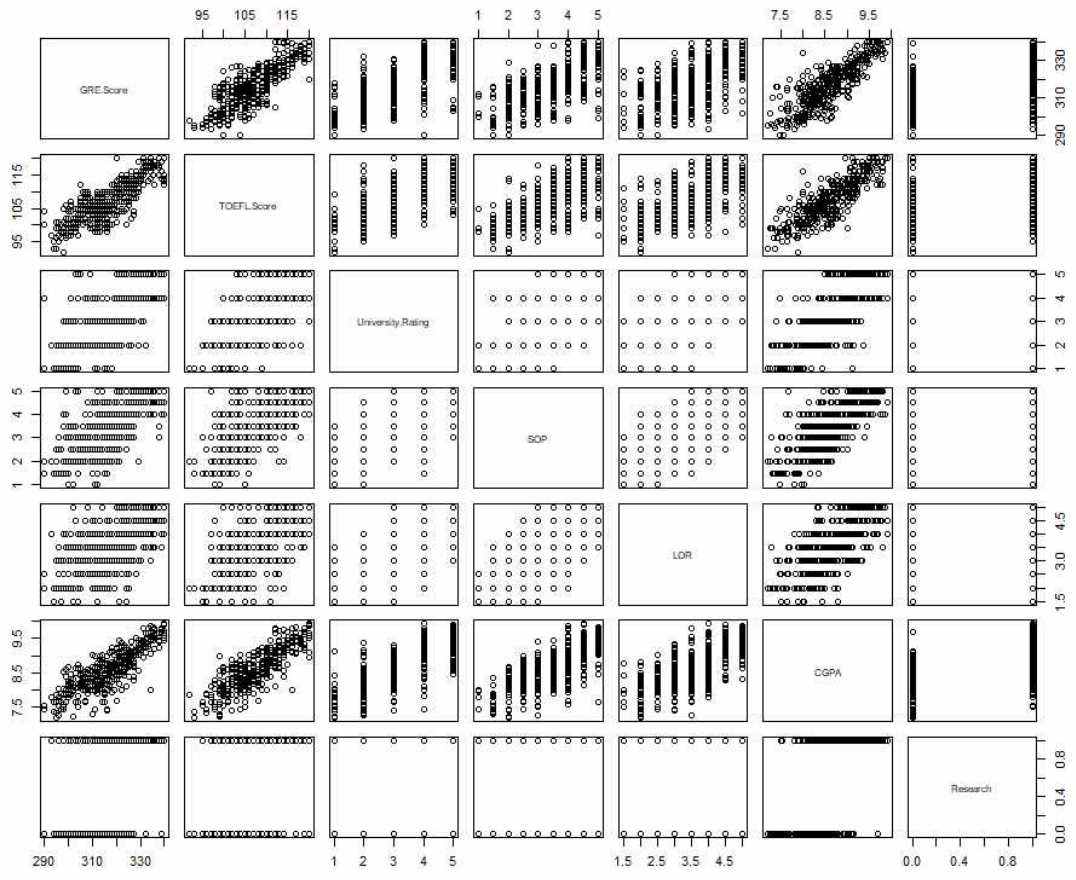
Kurtosis 값은 모두 3보다 작지만, Research를 제외한 변수들은 3에서 크게 벗어나지 않았으므로, 정규분포를 따른다고 가정하는 것이 타당하다.

[Q3]

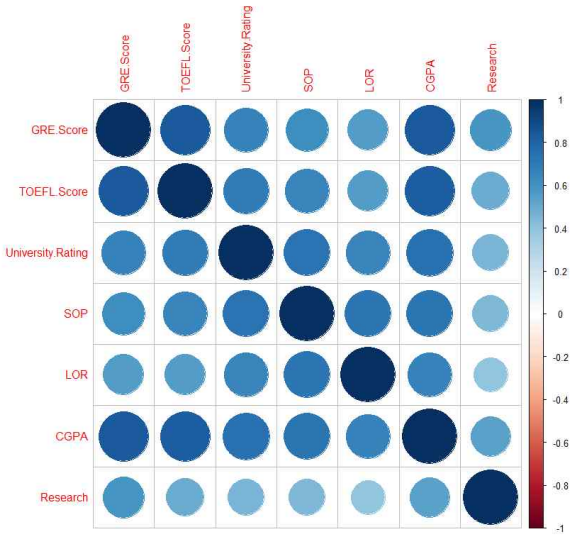
Q1에서 Q3 사이인 사분위간 범위(IQR)로 몸통을 구성하고 근접값들로 꼬리를 구성한다. 단위척도(step)는 1.5 x IQR이다. 안 울타리(inner fence)는 Q1에서 왼쪽으로 1 step만큼 간 것과 Q3에서 오른쪽으로 1 step만큼 간 것이다. 바깥 울타리(outer fence)는 Q1에서 왼쪽으로 2 step만큼 간 것과 Q3에서 오른쪽으로 2 step만큼 간 것이다. box plot을 기준으로 이상치는 whisker 범위보다 밖에 있으면 이상치로 정의하였다. 이에 따라 이상치를 제거하는 함수를 만들었고 이상치에 해당하는 객체들을 제거하였다. 이에 따라 원래 400개의 관측치에서 398개의 관측치를 가지는 데이터로 바뀌었다.

[Q4]

각 입력변수 간 산점도 결과는 다음과 같다.



corrplot 함수를 사용한 상관관계 시각화 결과는 다음과 같다.



산점도와 corrplot 함수의 결과를 분석해보았을 때, GRE.score와 CGPA가 가장 강한 상관관계를 보이는 것으로 나타난다. 이 외에 TOEFL.score와 GRE.score, TOEFL.score와 CGPA도 강한 상관관계를 보인다. 그도 그럴것이, GRE.score, TOEFL.score, CGPA 등 거의 모든 입력변수가 학업능력에 관한 변수이므로 변수들 사이에 강한 상관관계를 보이는 것이다.

[Q5]

입력변수에 대하여 정규화를 진행하였고, 종속변수를 Chance of Admit으로 두고 Logistic Regression Model을 적용하였을 때 결과는 다음과 같다.

```
Call:
glm(formula = Chance.of.Admit ~ ., family = binomial, data = Admission_trn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2753  -0.6466  -0.3355   0.5945   2.8171

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.5077     0.2035  -7.408 1.28e-13 ***
GRE.Score          -0.6703     0.3867  -1.733 0.083025 .
TOEFL.Score         0.1695     0.3305   0.513 0.608096
University.Rating   1.2698     0.2880   4.409 1.04e-05 ***
SOP                -0.9481     0.3276  -2.894 0.003804 **
LOR                 0.4597     0.2603   1.766 0.077324 .
CGPA                1.6329     0.4740   3.445 0.000571 ***
Research           -0.2523     0.2166  -1.165 0.244022
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 330.64  on 278  degrees of freedom
Residual deviance: 230.31  on 271  degrees of freedom
AIC: 246.31

Number of Fisher Scoring iterations: 5
```

유의수준 0.1에서 종속변수 Chance of Admit에 유의미하게 영향을 주는 변수는 University.Rating, SOP, CGPA가 있다.

[Q6]

Q5에서 training data set으로 설정한 Logistic Regression Model로 test data에 대한 예측을 진행하였다. cut-off value는 0.5로 설정하였을 때, confusion matrix와 True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, Simple Accuracy, Balanced Correction Rate, F1-Measure 값은 다음과 같다.

```
          lr_predicted
lr_target  0  1
        0 71 10
        1 14 24
                    TPR (Recall) Precision       TNR       ACC       BCR        F1
Logstic Regression    0.6315789 0.7058824 0.8765432 0.7983193 0.7440472 0.6666667
```

단순 정확도인 Accuracy 값은 0.79로 상당히 높은 수치를 보인다. BCR값은 0.74이고 F1-measure는 비교적 낮은 0.66의 수치를 보인다. confusion matrix를 통해 알 수 있는 사실은 Logistic Regression 모델이 0으로 예측했을 때 정확도가 1로 예측했을 때보다 비교적 크다는 것이다.

두 번째로는 cut-off value를 0.7로 설정해보았다.

```
          lr_predicted2
lr_target  0  1
        0 75  6
        1 23 15
>
> perf_mat[1,] <- perf_eval2(cm_full2)
> perf_mat
                    TPR (Recall) Precision       TNR       ACC       BCR        F1
Logstic Regression    0.3947368 0.7142857 0.9259259 0.7563025 0.6045635 0.5084746
```

cut-off value를 0.5로 설정했을 때와 비교해보면 accuracy, BCR, F1 값이 모두 다소 감소한 것을 볼 수 있다. Confusion matrix에서와 Recall 값에서 볼 수 있듯이 실제 값이 1일 때 예측을 굉장히 못하는 것을 볼 수 있다.

세 번째로는 cut-off value를 0.3으로 설정해보았다.

```
          lr_predicted3
lr_target  0  1
        0 64 17
        1  8 30
>
> perf_mat[1,] <- perf_eval2(cm_full3)
> perf_mat
                    TPR (Recall) Precision       TNR       ACC       BCR        F1
Logstic Regression    0.7894737 0.6382979 0.7901235 0.789916 0.7897985 0.7058824
```

cut-off value를 0.5로 설정했을 때와 비교해보면 accuracy값은 조금 감소했지만 BCR과 F1 값이 다소 상승한 것을 볼 수 있다. 이 경우 Recall 값도 상당히 상승한 것을 볼 수 있다. 이 경우는 data analyst가 관심있는 값(1), 이 경우에는 Chance of Admit이 종속변수 값들중에 minority인 것을 알 수 있다. 고로, cut off-value는 비교적 0.3이 적당한 값임을 알 수 있다.

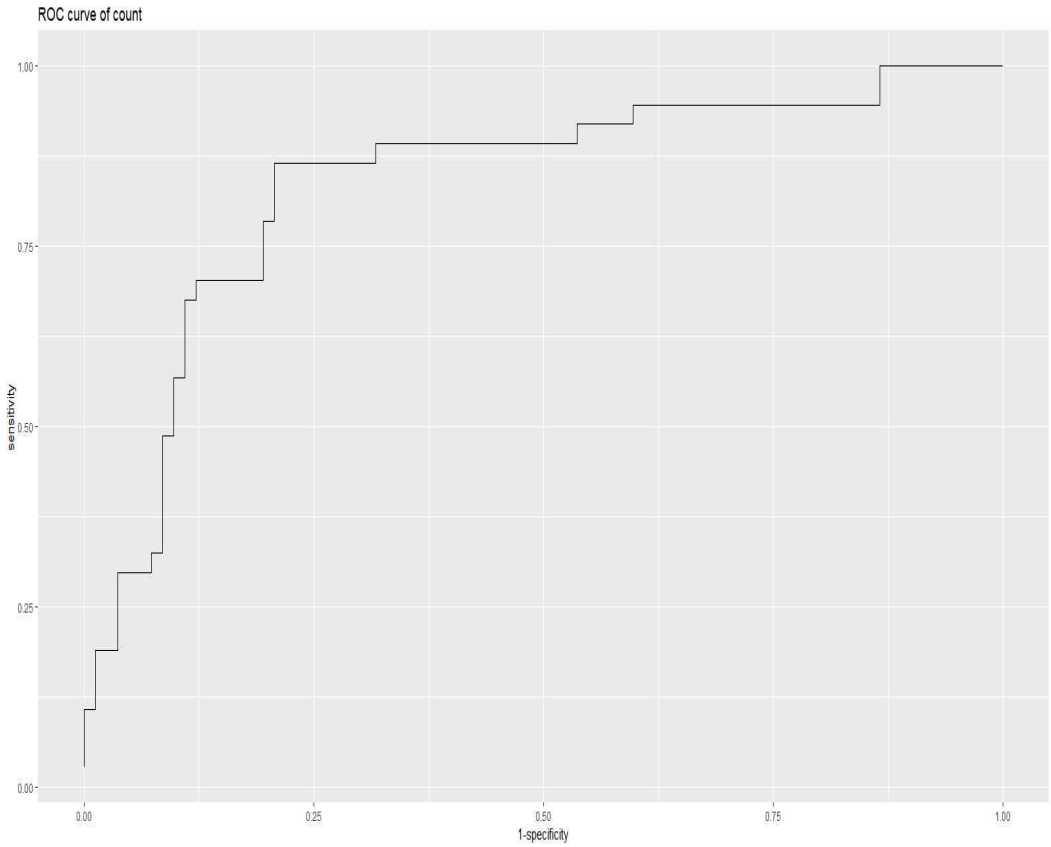
[Q7]

random seed값을 바꿔가면서 auroc 값을 구해봤을 때 다음과 같다.

```
      AUROC
[1,] 0.8306560
[2,] 0.8178788
[3,] 0.8520085
[4,] 0.8988724
[5,] 0.8375082
```

AUROC값은 대부분 0.8이상의 값을 가지므로 꽤 정확한 예측을 하고 있음을 나타낸다.

5번째 반복에 대한 결과에 대해 ROC curve를 시각화한 결과는 다음과 같다.



비교적 높은 넓이를 가진다는 것을 알 수 있다.

[Extra Question]

Extra Question으로는 kaggle에서 가장 유명한 datasets들중 하나인 credit card fraud를 로지스틱 회귀를 이용하여 분석을 진행해보았다. 이 datasets는 284,807건의 거래중 492건의 fraud 거래가 포함된 데이터이다. 입력변수는 30개를 포함하고 있고 그 중 28개의 입력변수는 PCA 변환이 된 변수이다. 나머지 두 개의 변수는 Time과 Amount이다. Time 변수는 거래간 시간이 얼마나 걸렸는지를 나타내는 변수이다. Amount는 거래량을 의미한다. 종속변수는 fraud 거래라면 1 아니면 0을 나타낸다.

데이터셋을 7:3으로 training dataset과 test dataset으로 나누어서 분석을 진행하였다. training dataset을 사용하여 logistic regression model을 구성한 결과는 다음과 같다.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.778e+00  3.129e-01 -28.052  < 2e-16 ***
Time        -1.409e-06  2.703e-06  -0.521 0.602170
V1           1.113e-01  4.941e-02   2.252 0.024338 *
V2          -4.948e-02  6.928e-02  -0.714 0.475127
V3           4.436e-02  6.557e-02   0.676 0.498757
V4           6.967e-01  9.330e-02   7.467 8.18e-14 ***
V5           7.000e-02  8.394e-02   0.834 0.404292
V6          -8.134e-02  9.014e-02  -0.902 0.366898
V7          -1.260e-01  7.908e-02  -1.593 0.111074
V8          -1.511e-01  3.400e-02  -4.445 8.80e-06 ***
V9          -4.287e-01  1.429e-01  -3.000 0.002703 **
V10         -9.844e-01  1.150e-01  -8.561  < 2e-16 ***
V11         -1.186e-01  9.978e-02  -1.188 0.234684
V12          3.955e-01  1.196e-01   3.306 0.000946 ***
V13         -4.750e-01  1.093e-01  -4.345 1.39e-05 ***
V14         -7.270e-01  8.254e-02  -8.808  < 2e-16 ***
V15          8.516e-03  1.042e-01   0.082 0.934893
V16         -4.161e-02  1.657e-01  -0.251 0.801729
V17         -8.606e-02  8.475e-02  -1.015 0.309894
V18         -8.815e-02  1.658e-01  -0.532 0.594863
V19          2.403e-01  1.220e-01   1.970 0.048838 *
V20         -4.629e-01  9.072e-02  -5.102 3.35e-07 ***
V21          4.355e-01  7.075e-02   6.155 7.49e-10 ***
V22          5.490e-01  1.585e-01   3.464 0.000533 ***
V23         -5.177e-02  6.813e-02  -0.760 0.447358
V24          1.670e-01  1.796e-01   0.930 0.352416
V25         -2.390e-03  1.569e-01  -0.015 0.987845
V26         -1.311e-01  2.249e-01  -0.583 0.559850
V27         -1.006e+00  1.255e-01  -8.016 1.09e-15 ***
V28         -3.885e-01  9.888e-02  -3.929 8.52e-05 ***
Amount       7.529e-04  3.998e-04   1.883 0.059687 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5039.2  on 199364  degrees of freedom
Residual deviance: 1539.5  on 199334  degrees of freedom
AIC: 1601.5
```

유의수준 0.1에서 15개의 입력변수가 유의미하다. 거의 모든 입력변수가 PCA 변환이 되어 있기 때문에 계수의 해석이 불가능하다.

Training set으로 구성한 logistic regression model로 test dataset을 예측해보았다. fraud transaction은 전체에서 굉장히 낮은 비율을 차지하기 때문에 cut-off value는 0.3으로 설정하였다. 그 결과 confusion matrix와 다양한 평가지표에 따른 수치는 다음과 같다.

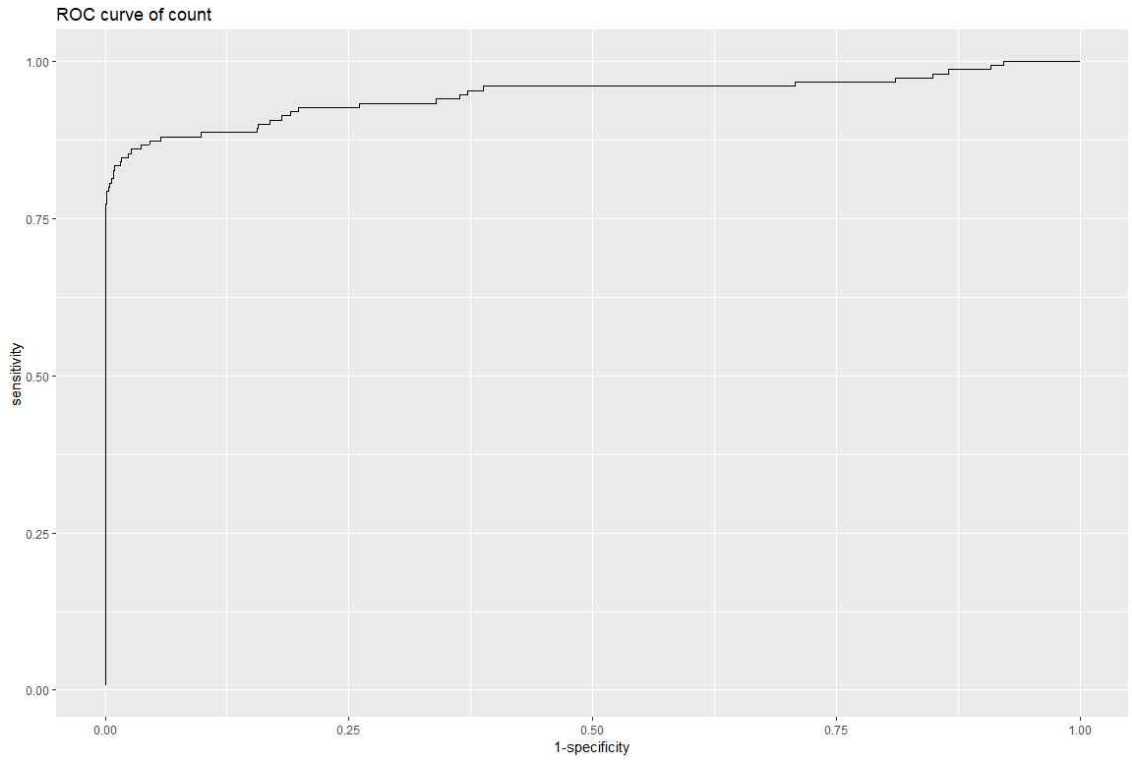
```
          lr_predicted1
lr_target     0     1
        0 85276    16
        1    48   102
                  TPR (Recall) Precision       TNR       ACC       BCR       F1
Logstic Regression        0.68 0.8644068 0.9998124 0.999251 0.8245438 0.761194
```

confusion matrix를 보면 실제 fraud transaction 150건 중 102건의 fraud detection에 성공한 것을 볼 수 있다. 이에 따른 Recall은 0.68로 꽤 높은 수치를 가진다. 다른 accuracy나 BCR, F1-Measure도 꽤 높은 값을 가진다.

다음으로는 AUROC값을 구해보았다.

```
> AUROC(roc_table)
[1] 0.9452545
```

AUROC값은 약 0.94로 굉장히 높은 값을 가진다. 1에 가까운 값이므로 모델링한 회귀식이 꽤 좋은 예측력을 가지고 있다고 봐도 무방하다.



ggplot을 활용해 ROC curve를 그려보았고, 굉장히 좋은 모습을 띄고 있다.