**주제: “Logistic Regression을 이용한 diabetes 예측”**

해당 자료는 본래 “National Institute of Diabetes and Digestive and Kidney Disease” 로부터 제공되었으며, diabetes “당뇨병”에 중점을 둔 자료이다. 데이터셋을 선정하게 된 이유는 여러 logistic regression과 가장 관련 있는 분야 중 하나가 건강 분야이며 여러 건강 관련 health 관련 데이터 셋들 중 당뇨에 대해 관심이 있기 때문이다. diabetes 데이터 셋은 총 768의 관측치와 9개의 변수들이 존재함으로 모델의 설명력이 좋을 것이라고 판단하였으며, 9개의 변수들이 모두 당뇨와 연관성이 높아 보이고 numerical인 점 또한 위 데이터를 선정하는데 큰 역할을 하였다. Diabetes 데이터 셋에 명시된 모든 환자들은 Pima Indian의 혈통이며, 최소 21살인 여성들이다. PIMA 여성들의 당뇨병 발병 확률이 상당히 높아 위 데이터 셋의 샘플로 선정되었다고 한다.

해당 데이터는 총 9개의 열(variables)이 존재하는데, 이 중 8개가 설명변수 이며 1개의 변수가 종속변수이다. 8개의 변수를 통해 종속변수가 발생할 확률을 Logistic Regression으로 예측하는데 주 목적을 둔다. 데이터셋의 변수들의 설명은 아래와 같다

|  |  |  |  |
| --- | --- | --- | --- |
| No | 변수 명 | 종속/설명 변수 | 설명 |
| 1 | Pregnancies | 설명변수 | 임신 횟수 |
| 2 | Glucose | 설명변수 | 혈당 |
| 3 | Blood Pressure (mm Hg) | 설명변수 | 혈압 |
| 4 | Skin Thickness (mm) | 설명변수 | 피부 두께 |
| 5 | Insulin (muU/ml) | 설명변수 | 인슐린 수치 |
| 6 | BMI (kg/m^2) | 설명변수 | BMI 수치 |
| 7 | Diabetes Pedigree Function | 설명변수 | 당뇨병 혈통 기능 |
| 8 | Age (Years) | 설명변수 | 나이 |
| 9 | Outcome | 종속변수 | 당뇨병 발생 여부 |

대한당뇨병병학회에 의하면 당뇨병의 주 원인은 아직 완벽하게 규명되어 있지 않다고 한다. 부모가 모두 당뇨병인 경우 자녀 역시 당뇨병에 걸릴 확률은 30% 이라고 한다. 그러나 유전적 요인이 당뇨병을 무조건적으로 야기하는 것은 아니고 나머지는 환경적 요인들이 대부분이라고 한다.

위 데이터 셋의 변수들 중 당뇨병 발병 여부와 가장 큰 상관관계를 갖을 것으로 예상되는 변수들은 Insulin, Age, 그리고Diabetes Pedigree이다.

1. Insulin: 당뇨병은 호르몬의 분비에 큰 영향을 받는다. 이 중 insulin은 여러 호르몬들 중 가장 큰 비중을 차지하는데, 인슐린이 체내의 클리코겐을 증가시키고, 혈당의 양을 조절하는 역할을 하기 때문이다. 실제 당뇨병 및 정신병 치료 시 인슐린을 체내에 투여한다고 알려진다. 따라서 우리 몸의 인슐린 수치와 당뇨병은 큰 관계가 있을 것으로 판단하였다.
2. Age: 대부분 당뇨병을 앓고 있는 환자들은 나이가 있는 40~50대이후의 중년층들이다. 실제 당뇨병은 연령이 높아질수록 발병률이 증가하기 때문에 높은 상관관계를 지닐 것으로 판단하였다.
3. Diabetes pedigree: 이 변수는 환자의 가족관계 중 당뇨병을 보유하는 사람들이 얼마나 있는지에 대한 정보를 제공한다. 앞서 말했듯이 당뇨병은 유전자와 매우 밀접한 관계를 갖는다. 따라서 위 설명변수가 diabetes의 발병 여부에 큰 영향을 끼칠 것으로 판단하였다.

해당 변수들 외에도 BMI도 큰 역할을 할 것이라 추정하였지만, BMI는 비만도(체내 체지방률)를 설명하는게 아니라 몸무게 대비 키의 관계를 의미하기에 배제하였다.

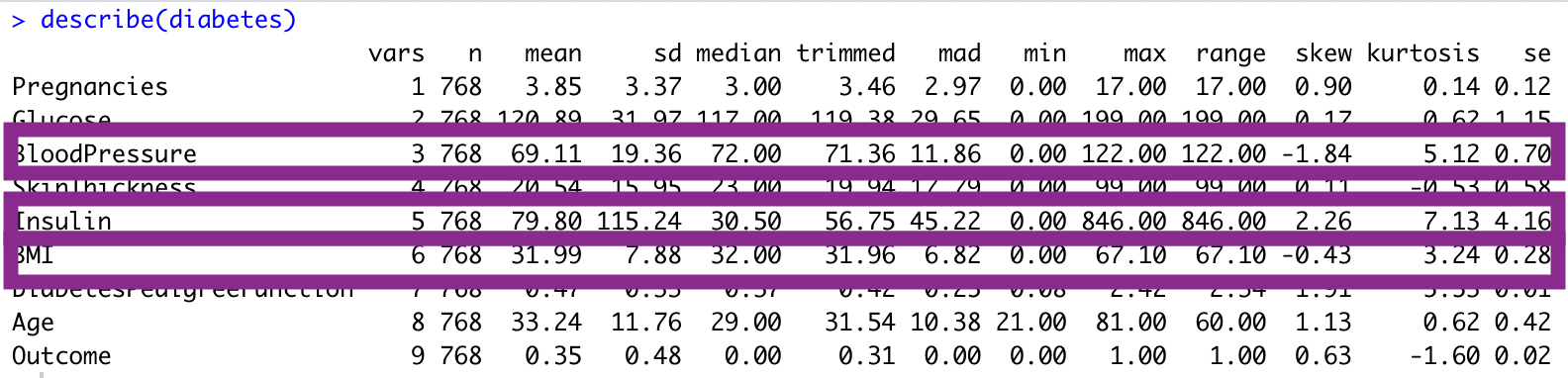
총 8개의 설명변수들 중 pregnancies, skin thickness 가 당뇨병에 대한 설명력이 가장 떨어지고 의미 없을 것으로 생각된다. 가장 큰 이유는 피부 두께와 임신 횟수가 당뇨병과 직접적인 관계를 갖는다는 것을 들어보지 못했다. 예를 들어 비만이거나 혈압이 높을수록 당뇨병이 있을 확률이 높다는 것은 누구나 대부분 아는 사실이나 피부 두께와 임신 횟수에 대한 언급은 들어본 적이 없기 때문이다. 대부분의 경우 당뇨병이 임신에 영향을 끼치는 경우는 있어도, 임신의 유무가 당뇨병에 영향은 끼치지 않는다고 한다. 또한 당뇨병이 있을 경우 피부의 색이 변하거나 건조하게 만들지만 피부두께에는 큰 변화가 나타나지 않는다고 한다.

**R-code**

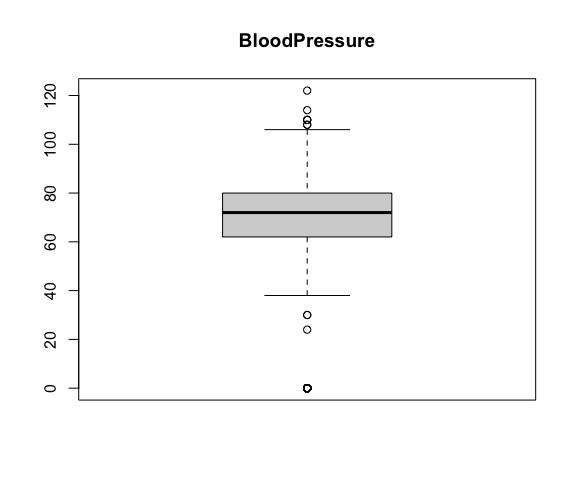
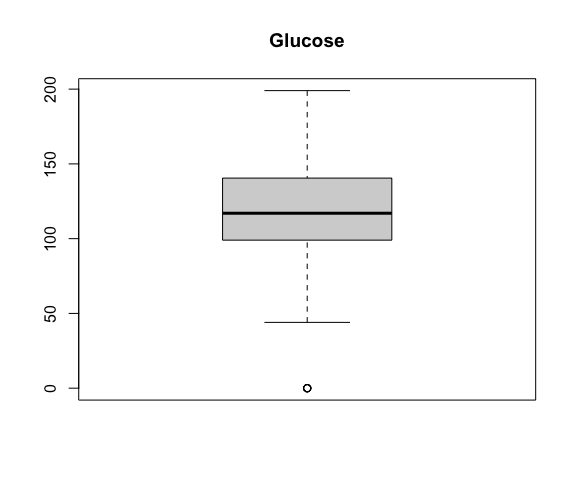
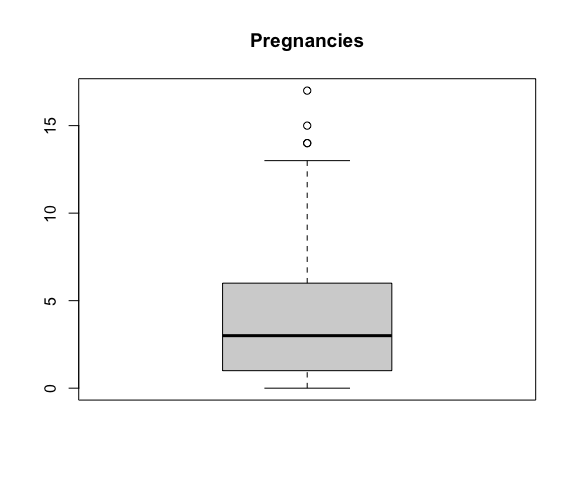
****

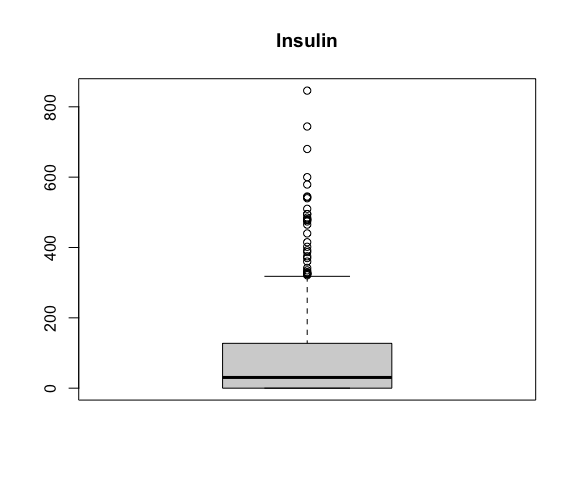
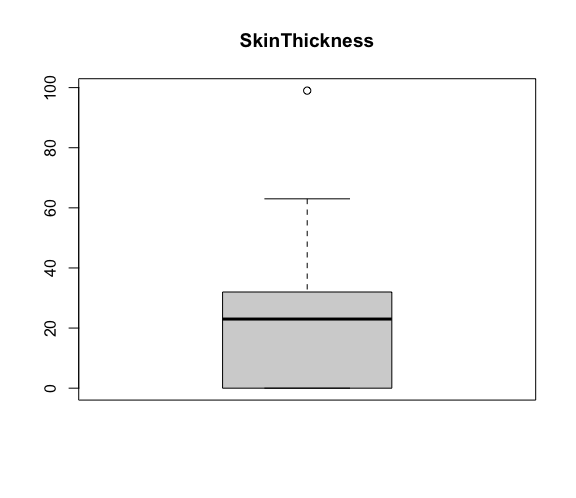
위 사진에 있는 R-code를 통해 단변량 통계량 및 Boxplot 그리고 평균, 표준편차, skewness, kurtosis를 계산하였다. 결과는 아래와 같다.

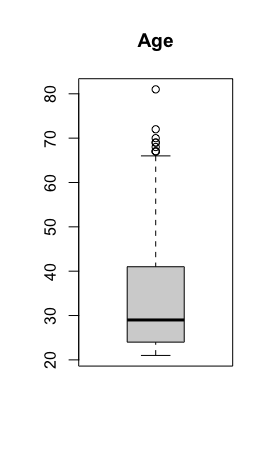
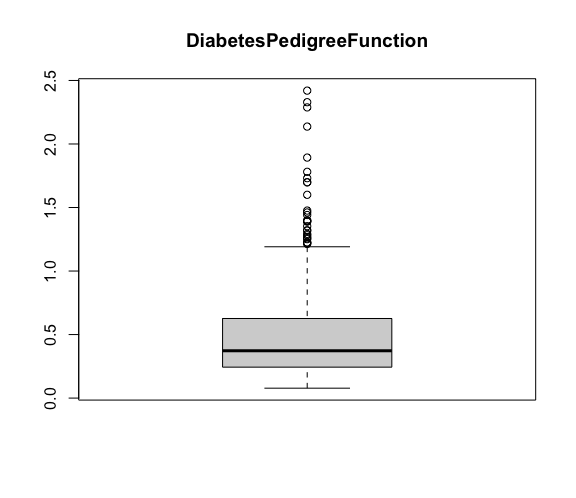
**Mean, Standard deviation, skewness, and kurtosis**

****

**Boxplots**







해당 변수들의 scale이 모두 다르기 때문에 다 다른 boxplot에 묘사를 하였다. 이상적인 정규분포는 skewness=0 그리고 kurtosis=3 가까운 값을 가져야 하며 boxplot은 symmetric하게 생길 수록 정규성을 갖는다. 그러나 사실 현실세계에서 skewness=0 그리고 Kurtosis 3인 데이터는 거의 존재하지 않기때문에 어느정도 감안을 해야한다. Boxplot을 보면 glucose, blood pressure, pregnancies,, diabetes pedigree function 그리고 bmi가 가장 symmetric 하게 생겼으며, describe 함수의 결과들 중 skewness와 kurtosis에 중점을 두고 본다면 blood pressure, insulin, 그리고 bmi가 정규성을 띈다. 결론적으로 Boxplot과 describe 함수의 결과 모두 높은 정규성을 보이는 blood pressure 와 bmi가 가장 높은 정규성을 띄고 glucose, pregnancies, diabetes pedigree function 그리고 insulin 역시 정규성을 갖는다.

위 boxplot을 보면, 8개의 설명변수들 중 총 2개의 변수가 다른 것들에 비해 이상치가 너무 큰 것을 볼 수 있다x. Insulin 과 Diabetes Pedigree Fucntion 둘다 maximum 값을 넘는 outlier들을 많이 갖는 것을 볼 수 있다. 아쉽게도 두 변수 모두 Q2에서 고른 종속변수와 상관관계가 높을 것으로 예상한 변수이지만 outlier가 워낙 높기에 데이터 셋에서 제거를 해준다. 해당 변수들의 인덱스 번호는 5,7이며 데이터 셋에서 제거하는 R-code와 새로운 만든 데이터 셋은 아래와 같다

**R-code**

텍스트이(가) 표시된 사진

자동 생성된 설명

**새로운 데이터셋**

테이블이(가) 표시된 사진

자동 생성된 설명

새로 만든 데이터 셋의 scatter plot과 correlation plot에 대한 R-code 및 결과는 아래와 같다. 여기서 diabetes\_data2 라는 데이터셋을 만든 이유는 diabetes\_data에서 7 번째 인덱스가 outcome 변수이며 이때 이는 factor형 이라서 correlation plot이 불가하기 때문이다.

**R-code**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**Scatter Plot**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**테이블이(가) 표시된 사진

자동 생성된 설명Correlation Plot**

**테이블이(가) 표시된 사진

자동 생성된 설명**

위에서 구한 correlation plot과 correlation coefficient 값들을 통해 어떤 변수들이 서로 강한상관관계를 갖는지를 알 수 있다. 주로 correlation coefficient의 절대값이 0.7이상이면 강한-상관관계, 0.5~0.7은 moderate correlation, 그리고 0.4 이하로는 약한 correlation으로 판단을 한다. 위 6개의 변수들 간의 상관관계 중 가장 높은 correlation coefficient를 갖는 것은 Age-Pregnancies 인데 아무래도 나이가 많을 수록 임신을 여러 번 할 기회가 많아서 그렇지 않나 싶다. 그러나 Age-Pregnancies의 상관계수는 0.57로 moderate correlation이며, 해당 데이터 셋의 변수들간의 상관관계가 그리 강하지 않음을 알 수 있다.

해당 모델의 변수가 이미 8개에서 6개로 줄었기 때문에, 아무래도 더 이상의 변수를 줄이는 것은 모델의 설명력을 많이 감소 시킬 것으로 판단한다. 더하여 해당 모델의 변수들 간의 강한 상관관계를 갖는 변수들은 없기때문에 데이터 셋에서 더 변수를 삭제할 필요는 없다.

해당 모델의 logistic regression을 위해 학습데이터와 테스트 데이터로 구분하여 분석을 실행한 코드는 아래와 같다.

**R-code텍스트이(가) 표시된 사진

자동 생성된 설명**

**로지스틱 분석 결과**

**테이블이(가) 표시된 사진

자동 생성된 설명**

위 로지스틱 회귀분석을 보면, ‘\*\*\*’인 변수들은 총 3개이며, 또한 ‘\*’인 변수는 한 개이다. 이때 ‘\*’의 개수는 해당 변수가 모델에 얼마나 유의미 한지 설명한다. 별의 개수가 많을수록 해당 모델에 대한 영향력이 크다 할 수 있다. 이때 별의 개수는 변수의 p-value 값에 의해 결정되는데 p-value 값이 낮다는 것은 귀무가설 (해당 변수가 영향력이 없다)을 reject하는 것이다. 또한 p-value 값이 0.05 미만일 경우부터 해당 변수가 유의미하다. 별이 총 3개인 변수는 Pregnancies, Glucose 그리고 BMI 이며, 그 다음으로 의미 있는 변수는 Age이다. 이때 총 4가지의 유의미한 변수들 중 Pregnancies는 당뇨병에 연관이 없다고 생각한다. 물론 당뇨와 임신이 영향을 가지기는 한다. 실제 미국에선 1%, 2%의 임산부들이 당뇨를 앓는다고 하며, 일반적으로 당뇨병 자체가 임신상태에 영향을 주지만 반대로 임신의 여부가 당뇨병에 영향을 끼치지는 않는다고 한다.

사람이 비만일 경우 일반적으로 당뇨병에 걸릴 확률이 더 높다. 물론 BMI 수치가 체질량 대비 키의 제곱으로 구해지기 때문에 해당 사람이 꼭 비만이란 판단을 내릴 수는 없다. 예를 들어 몸에 근육량이 많아 몸무게가 많이 나가는 경우는 비만이 아니기 때문이다. 그러나 BMI가 높을 수록 비만일 확률은 더 높다고 판단할 수 있기에 해당 모델에서 BMI가 유의미하게 측정 된 것 같다. Glucose의 경우 당뇨병과 직접적인 연관성을 갖는 것으로 알고있다. 우선 당뇨병 자체가 우리 몸이 에너지 생산을 위해 당을 쓰는 방식에 영향을 끼치는 질병이며, 당뇨병에 걸리면 포도당들이 혈류에 남아있고 세포들로는 전달이 되지 않아 세포들이 굶는다. 따라서 glucose가 높다는 것은 당뇨병에 걸릴 확률을 증가시킨다 생각한다. Age의 경우 대부분의 당뇨병 환자들이 중년층 들이기때문에 당연히 나이와의 연관성이 존재한다고 판단한다.

해당 모델에 영향력을 끼치지 않을 것으로 판단하였던 변수는 pregnancies와 skin thickness였다. Pregnancies의 경우 예상외로p-value값이 거의 0이며 해당 모델과 밀접한 관계를 갖는 것으로 분석이 되었다. Skin thickness의 p-value는 0.66이며 이는 상당히 높은 수치이다. 즉 Skin Thickness는 당뇨병의 발병 여부와는 연관성이 없다는 것이다.

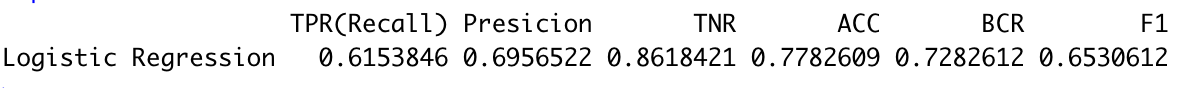
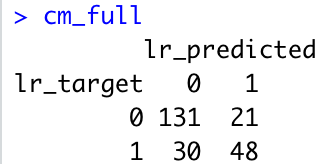
위 데이터 셋의 confusion matrix 생성 및 evaluation을 위해 아래 R-code를 활용하였다

**R-code**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

**Evaluation Measure**

** **

R-code를 통해 위와 같은 결과를 산출할 수 있었다. Confusion Matrix를 보면 해당 모델이 총 161개의 데이터가 당뇨병이 없다고 판단하여 131개를 옳게 판단하고 30개를 잘못 판단하였다. 반대로 총 69개를 당뇨병이 있다고 판단하였으나 이 중 21개는 실제 당뇨병이 없는 것으로 측정되었다.

좌측 evaluation 표를 살펴보면:

\*순수 정확도 (Accuracy)가 0.77이며 이때 전체 환자들 중 77%를 당뇨병이 있는지 없는지를 옳게 판단한 것이다.

\*TPR=0.61인데 해당 값은 실제 총 78명을 당뇨병 환자 중 모델이 해당 환자가 당뇨병을 앓고 있다고 판단하였을 때 확률을 보여준다 (48/78).

\*반대로 TNR은 실제 당뇨병이 없는 환자를 모델이 얼마나 잘 예측하는지를 나타낸다..

\*Precision 수치는 모델이 당뇨병이 있을 거라고 예측했을 때 실제 데이터와 비교한 정확도를 보여준다. (48/69)

\*BCR=0.728은 모델의 예측력을 각각 당뇨병이 있다/없다 categorical하게 판단한 확률을 기하평균을 사용하여 보여준다.

\*마지막으로 F1 Measure=0.653은 해당 모델의 Recall과 precision에 대한 기하평균 확률이라고 할 수 있다.

아래와 같은 코드를 통해 테스트 데이터 셋에 대한 AUROC 계산을 위한 함수를 생성하였다.

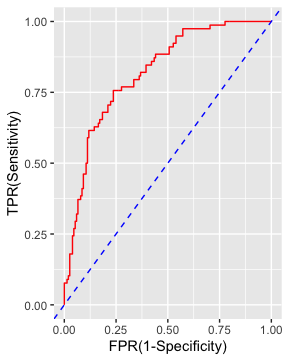
**R-code(Test Data)**



위 코드를 실행하여 ROC curve와 AUROC를 계산할 수 있었다. 결과는 아래와 같다.

**AUROC(Test Data) ROC Curve(Test Data)**

텍스트이(가) 표시된 사진

자동 생성된 설명 

Receiver operating characteristics, ROC는 분류 모델의 본질적인 성능에 대한 평가를 해준다. 이때 ROC curve는 true positive rate TPR과 false positive rate 이 두가지 지표를 사용하여 표현이 된다. 해당 모델에서 TPR은 실제 당뇨병이 있는 사람이 당뇨병이 있다고 예측될 확률이며, FPR은 실제 당뇨병이 없는 사람이 당뇨병이 없다고 판단될 확률이다. 해당 모델의 ROC그래프는 좌측 그래프와 같이 표현되었다. 이때 파란색 선은 Random classifier의 ROC 커브이며, 빨간색 선이 테스트-데이터셋의 ROC curve이다. AUROC 즉 ROC curve 아래 영역의 넓이를 나타내는 지표는 0.5에서부터 1까지 값을 갖는다, 이때 좋은/이상적인 분류 모델은 ROC 선 아래의 영역이 1이며, ROC curve가 좌상단에 가까울 수록 좋은 모델이다. 테스트 데이터에 대한 분류 모델의 AUROC(ROC curve의 넓이)는 0.8188로 상당히 높은 수치가 측정된 것을 볼 수 있다. 즉 당뇨병 분류 모델이 테스트 데이터를 분류하는데 높은 성능을 갖고 있음을 의미한다.

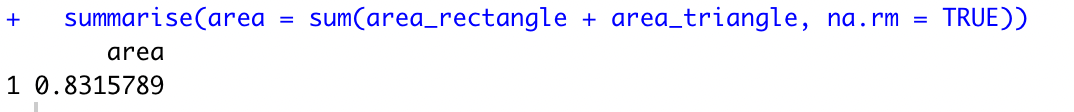
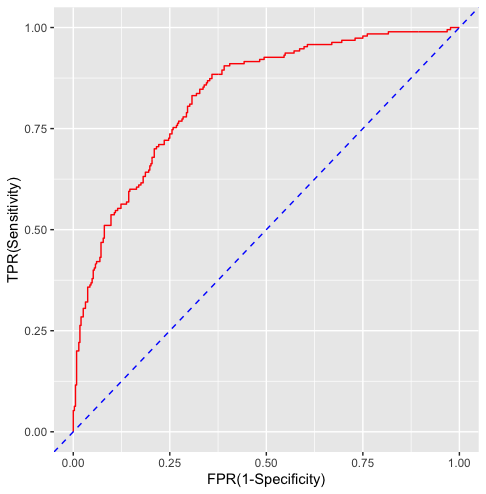
아래와 같은 코드를 통해 학습 데이터 셋에 대한 AUROC 계산을 위한 함수를 생성하였다.

**R-code(Test Data)**

텍스트이(가) 표시된 사진

자동 생성된 설명

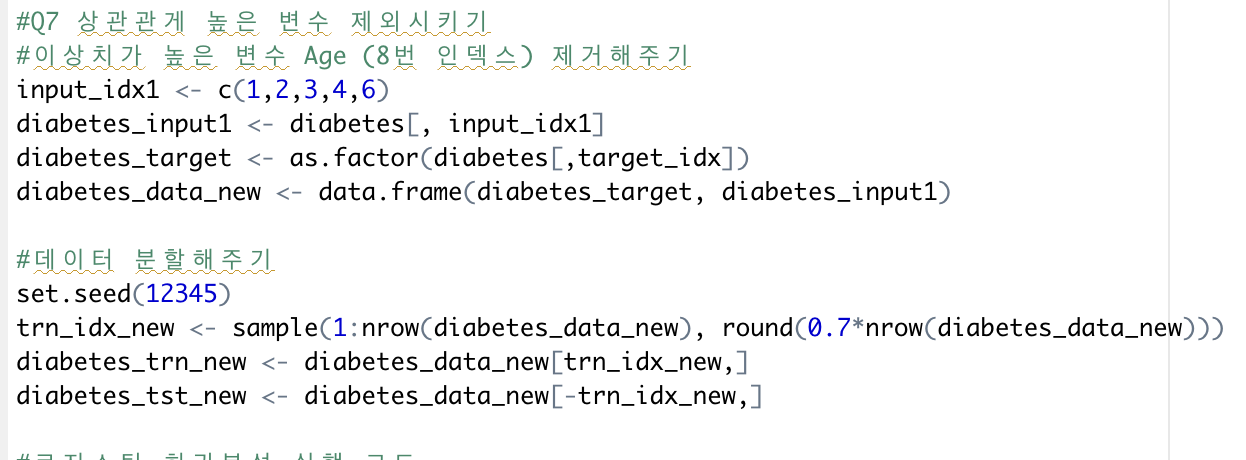
**AUROC(Train Data) ROC Curve(Train Data**

** **

위 AUROC 값과 ROC-Curve는 학습데이터에 대한 것이다. 위에서 구한 테스트 데이터 셋과의 ROC curve를 비교하였을 때 학습데이터의 ROC curve가 더 좌상단에 가까우며 넓이 역시 0.8315로 테스트 데이터 셋의 AUROC보다 더 크다. 학습 데이터의 AUROC가 더 크기 때문에 해당 로지스틱 회귀분석은 학습데이터를 예측하는데 더 좋음을 알 수 있다. 이는 모델 자체를 학습시킬 때 학습데이터를 기반하였기 때문에 당연히 학습데이터에 대한 예측률이 더 좋다고 할 수 있다.

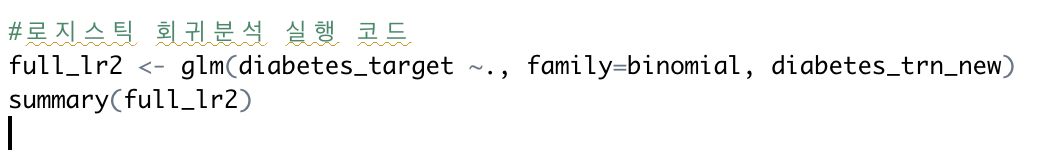
이전 단계의 상관관계 구할 때, 유일하게 significant한 상관관계를 보인 것은 Age-Pregnancies였다. 데이터 셋의 변수들은 정규성 검정 후 종속변수를 제외하고 총 6개의 변수들 만이 남아있어, 상당히 변수가 적은 편에 속한다. 따라서 Age와 Pregnancies 중 p-value값이 더 낮게 측정되었던 Age를 제외하고 logistic regression 모델 학습을 진행하였다. 해당 모델의 데이터셋을 새로 편성하는 코드는 아래와 같다.

**R-code**



변수들의 유의수준을 계산하기 위해 아래와 같은 R-code를 사용하였다.

**R-code**



**테이블이(가) 표시된 사진

자동 생성된 설명** 테이블이(가) 표시된 사진

자동 생성된 설명오른쪽에 보이는 결과가 현재 새로 구축해 사용하는 데이터에 대한 summary함수의 적용 결과이고, 좌측의 결과는 이전 단계에서 사용한 데이터에 관한 것이다. 이때 deviance residual을 보면, Age를 제외한 후 logistic model의 분포도가 커진 것을 볼 수 있다. 예를 들어 min 값이 -2.205에서 -2.27로 증가하였고, max값 역시 2.7441에서 2.838로 증가하였다. 이는 변수가 하나 줄어들면서 나온 당연한 결과라 볼 수 있다. 변수를 삭제 후에도 Pregnancies, Glucose 그리고 Bmi 수치는 여전히 모델에 유의미한 영향을 끼치고 있으며 Skinthickness의 경우 p-value값이 감소하였고, BloodPressure은 증가한 것을 볼 수 있다.

Confusion Matrix 생성을 위해 아래 코드를 사용하였다

**R-code**

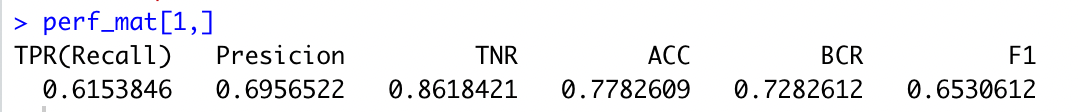
**텍스트이(가) 표시된 사진

자동 생성된 설명이전 데이터 셋. 현재 데이터 셋의**

**테이블이(가) 표시된 사진

자동 생성된 설명 테이블이(가) 표시된 사진

자동 생성된 설명**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

TPR을 제외한 모든 evaluation 수치들이 Age를 데이터 셋에서 제거 후 감소함을 볼 수 있다. 이는 Age가 상관관계를 갖더라도 p-value 0.02임으로 해당 모델에 그래도 유의미한 변수이기 때문이다. 즉 Age가 삭제되면서 해당 모델의 예측력이 전반적으로 감소했다고 할 수 있다.

학습/테스트 데이터 셋을 위한 ROC\_Curve 와 AUROC를 계산하기 위해 아래와 같은 R\_code를 활용하였다.

**R-code**

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

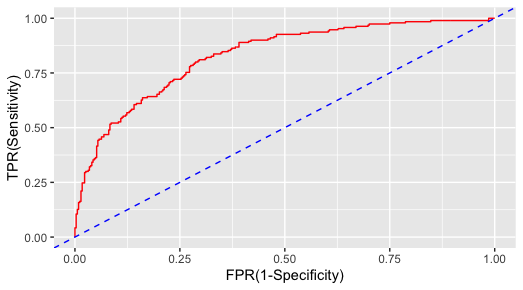
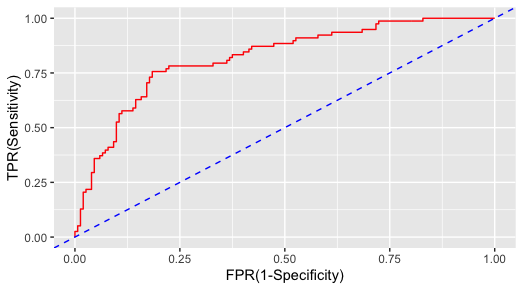
자동 생성된 설명

Test -Data Set. Train-Data Set

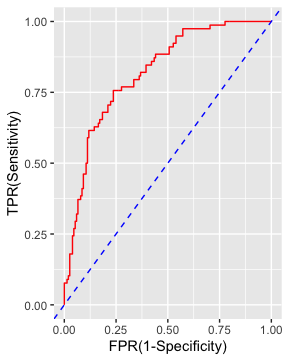
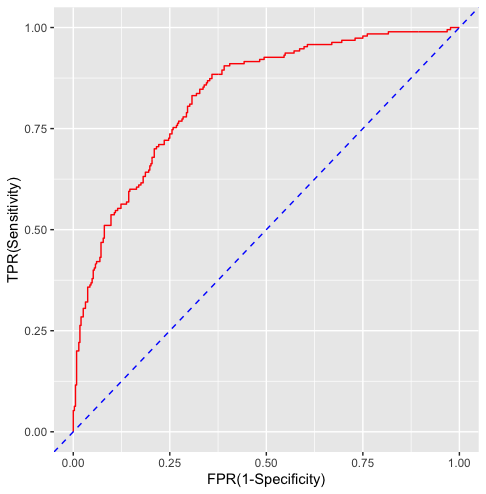
텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명



Old-Test -Data Set. Old Train-Data Set

****

위 Q[7]의 ROC 커브를 보면 모두 좌상향쪽으로 치우친 것을 볼 수 있다.

위 결과를 통해 Q[6]의 결과들과 비교를 해보면 ROC curve를 통해서는 그닥 그리 큰 차이를 볼 수는 없다. 그렇기 때문에 수치화된 AUROC로 성능을 비교하기로 하였다.

|  |  |  |  |
| --- | --- | --- | --- |
| New Test | New Train | Old Test | Old Train |
| 0.819 | 0.828 | 0.,811 | 0.831 |

New AUROC 모두 0.8에 가까운 수치이며 이는 어느정도 준수한 수치라고 할 수 있다. Old 와 비교를 하였을 때 크게 차이가 나지않으며, test에 대한 AUROC는 증가하였고 Train에 대한 AUROC는 감소 했음을 볼 수 있다. 전반적으로 Age가 빠진 후에도 logistic regression의 성능이 크게 떨어지지 않았다.

|  |
| --- |
|  |

선형회귀는 least square estimation을 통해 R제곱을 구할 수 있지만, 반대로 달리 logistic regression에선 R제곱 통계량을 통해 설명변수들의 분산의 비율을 설명을 할 방법이 없다. 따라서 logistic 회귀분석의 경우 pseudo R제곱을 구한다. 이 분석에서 가장 중요한 것은 McFadden’s R2 값인데, 아래와 같은 식으로 계산된다.

🡪 ln(LM) 값은 fitted model에 대한 Log likelihood

🡪 ln(L0) 값은 null model에 대한 log likelihood

여기서 null model이란 random classifier를 의미한다. 즉 우리가 예측하고자 하는 종속변수 y 값에 대한 예측을 변수들의 사용과 근거 없이 그냥 random하게 예측하는 것이다. 즉 위McFadden식은 구현한 모델과 random classifier의 예측력의 비율을 1에서 빼준 값이다. .McFadden’s R2는 0에서1까지의 값을 가지며, 0.2에서 0.4사이의 값이 좋은 예측력을 가졌다고 할 수 있다.

Diabetes 데이터 셋에 대한 예측력을 evaluate 하기 위해 McFadden을 구현하는 R-code는 아래와 같다.

텍스트이(가) 표시된 사진

자동 생성된 설명 텍스트이(가) 표시된 사진

자동 생성된 설명

위 pR2코드를 실행하여 diabetes 로지스틱 회귀분석 모델에 대한 McFadden값은 0.2644로 측정이 되었다. 앞서 말했듯 excellent fit은 0.2와 0.4사이의 Pseudo R2값을 갖기 때문에 diabetes logistic regression은 좋은 모델이라고 결론을 내릴 수 있겠다. 실제 위에서 구한 AUROC값 역시 0.8로 상당히 높은 수치였기 때문에 이러한 위와 같은 근거들로 해당 모델의 예측력이 좋은 편이라고 할 수 있다.

**당뇨병 관련 자료를 위해 사용한 참고 사이트**

<https://www.aad.org/public/diseases/a-z/diabetes-warning-signs>

https://www.diabetes.or.kr/general/class/index.php?idx=2

https://www.bd.com/ko-kr/offerings/diabetes-care/effective-diabetes-management/introduction-to-diabetes