머신러닝 기반 데이터 분석

XGBOOST 와 로지스틱회귀분석을 활용한 위스콘산 유방암 분석 예측 및 비교

2023.03.27

B2 팀

서영석, 박용태, 이현호, 전국림

<목차>

1. 서론

1) 데이터 분석 배경

이 데이터셋은 위스콘신 대학교 병원에서 수집된 Fine Needle Aspiration (FNA) 검사 결과를 기반으로 구축되었습니다. FNA 검사는 바늘을 사용하여 유방종양으로부터 세포 샘플을 추출하고 현미경으로 살펴보는 검사입니다. 이를 통해 유방종양의 양성/악성 여부를 판단할 수 있습니다.

569 개의 관측치와 30 개의 변수로 구성되어 있으며, 각각의 관측치는 유방조직 샘플에 대한 여러 가지 측정값을 나타냅니다. 이 중 212 개의 관측치는 악성(malignant) 종양을, 357 개의 관측치는 양성(benign) 종양을 나타냅니다.

각 관측치에 대해 측정된 30 개의 변수는 다양한 특징을 나타내며, 이러한 변수들을 사용하여 유방암 종양을 악성과 양성으로 분류하는 머신 러닝 모델을 만들 수 있습니다. 이 데이터셋은 머신 러닝 분류 모델 개발 및 평가를 위한 대표적인 데이터셋 중 하나입니다.

2) 데이터 분석 설명

이 보고서에서는 R 을 활용하여 유방암의 Benign(양성), Malignancy(악성) 판독이 각 관측치의 상관관계에 대한 데이터를 분석할 예정이다.

사용 분류기법 → xgboost , 로지스틱 회귀분석..

2. 본론

1) 데이터 로딩 및 탐색

먼저 위스콘신 유방암 데이터셋을 로딩하고.

```
wd_df <- read.csv("wdbc_data.csv", header = T)
str(wd_df)
```

2) 데이터 구조

데이터셋 내의 변수명의 의미를 아래 표와 같이 정리하였다

변수	변수 정의
diagnosis	유방종양의 진단결과 Benign(양성) 'B', Malignancy(악성) 'M'
radius_mean	종양 경계면에서 종양의 중심까지 거리의 평균값
texture_mean	그레이스케일 이미지의 질감(texture) 평균값
perimeter_mean	종양 경계면의 길이 평균값
area_mean	종양 영역의 면적 평균값
smoothness_mean	경계면의 국소적인 길이 차이의 제곱합에 대한 측정값
compactness_mean	종양의 평균 크기에 대한 경계면의 길이 제곱값과 종양 영역의 면적 비율의 비율
concavity_mean	종양 경계면에서 볼록한 부분의 수량 평균값
points_mean	볼록한 부분의 선분의 길이 합계 평균값
symmetry_mean	두 부분 사이의 균형 정도의 측정값
dimension_mean	종양의 모양을 설명하는 척도로, 세 가지 주요한 축의 길이에 대한 평균값
radius_se	경계면에서 종양 중심까지의 거리에 대한 표준오차값
texture_se	그레이스케일 이미지의 질감에 대한 표준오차값
perimeter_se	종양 경계면의 길이에 대한 표준오차값
area_se	종양 영역의 면적에 대한 표준오차값
smoothness_se	경계면의 국소적인 길이 차이의 제곱합에 대한 표준오차값
compactness_se	경계면의 길이 제곱값과 면적 비율의 비율에 대한 표준오차값
concavity_se	종양 경계면에서 볼록한 부분의 수량에 대한 표준오차값
points_se	볼록한 부분의 선분의 길이 합계에 대한 표준오차값
symmetry_se	두 부분 사이의 균
dimension_se	세 가지 주요한 축의 길이에 대한 표준오차값
radius_worst	종양 경계면에서 종양의 중심까지 거리의 최대값
texture_worst	그레이스케일 이미지의 질감 최대값
perimeter_worst	종양 경계면의 길이 최대값
area_worst	종양 영역의 면적 최대값
smoothness_worst	경계면의 국소적인 길이 차이의 제곱합에 대한 최대값
compactness_worst	종양의 평균 크기에 대한 경계면의 길이 제곱값과종양영역의 면적 비율의 비율에 대한 최대값
concavity_worst	종양 경계면에서 볼록한 부분의 수량 최대값
points_worst	볼록한 부분의 선분의 길이 합계 최대값
symmetry_worst	두 부분 사이의 균형 정도의 측정값의 최대값
dimension_worst	종양의 모양을 설명하는 척도로, 세 가지 주요한 축의 길이에 대한 최대값

.총 30 개의 변수로 구성되였고 569 관측치로 구성되어 있다

3) 분석을 위한 데이터 전처리

```
#종속변수는 diagnosis: Benign(양성), Malignancy(악성)
#양성이면 0 악성이면 1 그외 2 로 정제
wd_label <- ifelse(wd_df$diagnosis == "B", 0,
                   ifelse(wd_df$diagnosis == "M", 1, 2))
#종속변수 라벨링
wd_df$label <- wd_label
#트레이닝, 테스트 데이터 셋 생성
set.seed(1234)
idx <- sample(nrow(wd_df), 0.7 * nrow(wd_df))
#학습 데이터 70%
train <- wd_df[idx,]
#테스트 데이터 30%
test <- wd_df[-idx,]
```

분석을 위하여 종속변수인 **diagnosis 변수 데이터 기준으로** 양성이면 0 악성이면 1 그 외는 2 로 정제하여 새로운 변수 label 에 저장한다

4-1) 데이터 분석(xgboost)

-xgboost 데이터 전처리

```
library(xgboost)
#matrix 객체 변환
train_mat <- as.matrix(train[-c(1, 31)])
train_lab <- train$label
#test set 생성
test_mat <- as.matrix(test[-c(1, 31)])
test_lab <- test$label
```

xgboost 함수를 사용하기 위하여 매트릭스 형태로 변환이 필요하다. as.matrix()함수를 통해 종속변수를 제외한 나머지 변수들을 **train_mat** 에 저장하고 종속변수는 **train_lab** 에 저장한다.

- xgboost 모델 생성

```
#xgb.DMatrix 객체 변환
dtrain <- xgb.DMatrix(data = train_mat, label = train_lab)
#xgb model 생성
xgb_model <- xgboost(data = dtrain, max_depth = 2, eta = 1,
                     nthread = 2, nrounds = 2, num_class = 2,
                     objective = "multi:softmax",
                     verbose = 0)
```

트리의 최대 깊이인 max.depth 2, 시행속도 eta 는 1 로 XGBoost 를 실행하는 데 사용되는 병렬 스레드 수 nthread 는 2, 최종 모델의 결정 트리 수 nrounds 는 2, "multi:softmax"는 다중 클래스 분류를 수행하도록 XGBoost 를 설정한다.

- xgboost 분류 모델 평가

```
#모델 예측치
pred_wd <- predict(xgb_model, test_mat)
table(pred_wd, test_lab)
#정확도 확인
sum(pred_wd == test_lab)/ NROW(pred_wd) # (110+48) /171 =0.92239766
# 정확성 92%
```

코드 실행결과:

	Test_lab(실제값)	
pred_wd(예측값)	0	1
0	**110**	11
1	2	**48**

이에 해당 분류모델의 **Accuracy 정확도는 (110+48)/171 = 0.92397 약 92%이다**

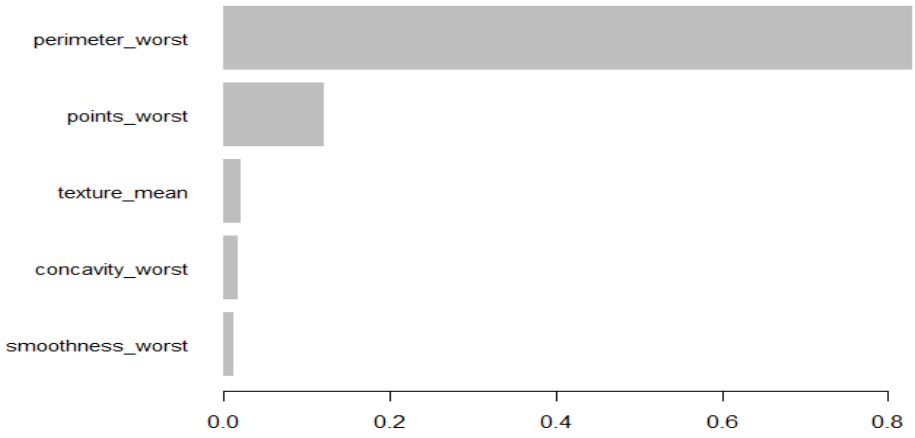
- xgboost 분류 모델의 중요변수의 영향력 및 시각화

```
#변수 중요도
importance_matrix<-xgb.importance(colnames(train_mat), model = xgb_model)
importance_matrix
#중요 변수 시각화
xgb.plot.importance(importance_matrix)
```

실행결과:

Feature(변수)	Gain(기여도)	Cover(특징지수)	Frequency(사용빈도)
perimeter_worst	**0.82974383**	**0.50000000**	**0.3333333**
points_worst	**0.12080276**	**0.22835643**	**0.1666667**
texture_mean	**0.02076423**	**0.07057131**	**0.1666667**
concavity_worst	**0.01757083**	**0.10950922**	**0.1666667**
smoothness_worst	**0.01111835**	**0.09156304**	**0.1666667**

시각화:



- xgboost 변수 영향력 실행 결과 분석

perimeter_worst,points_worst,texture_mean,concavity_worst,smoothness_worst 5 개의 변수가 중요변수로 확인되고 그중 perimeter_worst(종양경계면의 길이 최대값)가 가장 큰 영향력을 가지고 있다.

본 분석을 통하여 유방암의 양성과 악성의 판독은 종양의 경계면(perimeter_worst)의 길이가 커질수록 악성이 확률이 높을것을 알수 있다. 또한 볼록한 부분의 길이,질감 등 수치도 커질수록 악성일 확률을 조금씩 증가하고 있다.

4-2) 데이터 분석(로지스틱회귀분석)

-로지스틱 회귀분석 데이터 전처리 및 모델 생성

```
library(car)
library(lmtest)
#결측치 확인
table(is.na(wd_df))
#훈련데이터로 1 차 회귀분석 모델 생성
wd_model <- glm(label ~ ., data = train)
#분석결과 확인
summary(wd_model)
```

실행결과:

> summary(wd_model)

glm(formula = label ~ ., data = train)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.48159	-0.15626	-0.01978	0.13740	0.85497

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.46166	0.554467	-2.636	0.00874
radius_mean	-0.39963	0.205958	-1.94	0.0531
texture_mean	-0.00083	0.009541	-0.087	0.93064
perimeter_mean	0.02476	0.028964	0.855	0.39318
area_mean	0.001632	0.000705	2.316	0.02113
smoothness_mean	0.658632	2.370032	0.278	0.78125
compactness_mean	-3.75995	1.576436	-2.385	0.01758
concavity_mean	1.344718	1.231479	1.092	0.27557
points_mean	2.975808	2.327308	1.279	0.20183
symmetry_mean	0.560954	0.917587	0.611	0.54136
dimension_mean	-0.78268	6.73278	-0.116	0.90752
radius_se	0.098117	0.363377	0.27	0.7873
texture_se	0.009268	0.044881	0.207	0.83651
perimeter_se	-0.02178	0.048901	-0.445	0.65628
area_se	-0.00031	0.001994	-0.153	0.87841
smoothness_se	18.82808	7.494857	2.512	0.01243
compactness_se	-1.02174	2.487807	-0.411	0.68153
concavity_se	-3.4876	1.47115	-2.371	0.01827
points_se	10.37695	6.867557	1.511	0.13165
symmetry_se	2.591617	3.574256	0.725	0.46887
dimension_se	0.030771	13.18751	0.002	0.99814
radius_worst	0.330874	0.073377	4.509	8.77E-06
texture_worst	0.006978	0.008308	0.84	0.40153
perimeter_worst	-0.00365	0.007141	-0.511	0.6099
area_worst	-0.00179	0.000445	-4.031	6.76E-05

smoothness_worst	-0.65254	1.630334	-0.4	0.68921
compactness_worst	0.313782	0.435186	0.721	0.47135
concavity_worst	0.325158	0.304925	1.066	0.28696
points_worst	0.258572	1.106483	0.234	0.81536
symmetry_worst	0.199383	0.607031	0.328	0.74275
dimension_worst	3.012395	2.772868	1.086	0.27802

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.05612246)

Null deviance: 94.183 on 397 degrees of freedom

Residual deviance: 20.597 on 367 degrees of freedom

AIC: 14.874

Number of Fisher Scoring iterations: 2

1 차 훈련데이터로 회귀분석한 결과 p-value 값이 0.05 보다 큰 변수들이 많을것을 알수 있다. 후진제거법으로 데이터 정제가 필요하다

-로지스틱 회귀분석 후진제거법 실행

```
# p-value 값이 0.05 보다 큰 유의않지않는 변수 제거(후진제거법)
wd_model2 <- step(wd_model, direction = "backward")
#추출된 변수 확인
formula(wd_model2)
summary(wd_model2)
```

실행결과:

```
> summary(wd_model2)
Call:
glm(formula = label ~ radius_mean + area_mean + compactness_mean +
    points_mean + smoothness_se + concavity_se + points_se +
    symmetry_se + radius_worst + texture_worst + area_worst +
    concavity_worst + dimension_worst, data = train)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.47524	**-0.15592**	**-0.02096**	**0.13761**	**0.83899**

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.59037	0.2732606	-5.82	1.24E-08

radius_mean	-0.215	0.0603402	-3.563	0.000413
area_mean	0.001385	0.0005262	2.631	0.008853
compactness_mean	-2.78215	0.6694103	-4.156	3.99E-05
points_mean	5.397696	1.0998477	4.908	1.36E-06
smoothness_se	16.18545	5.0615539	3.198	0.0015
concavity_se	-2.95115	0.826511	-3.571	0.000401
points_se	6.095983	3.942882	1.546	0.122911
symmetry_se	3.902593	1.6885971	2.311	0.021354
radius_worst	0.298797	0.0452244	6.607	1.31E-10
texture_worst	0.007318	0.0021754	3.364	0.000846
area_worst	-0.00175	0.0003269	-5.363	1.42E-07
concavity_worst	0.624746	0.1601037	3.902	0.000113
dimension_worst	3.253435	1.34437	2.42	0.015982

실행후 절편이 -1.5903718 인 모델이 생성된다. 즉 추출된 변수들의 값들이 증가 될수록 악성 판정 확률이 줄어듬을 알수 있다. 다음 모델로 데이터 검증으로 정확도 측정을 진행.

-로지스틱 회귀분석 모델 평가

```
#예측치 구하기
pred_wd2 <- predict(wd_model2, newdata =test, type = "response")
pred_wd2
# 예측치 컷오프를 0.5 로 가정하여 분류값을 생성한다. 0.5 이상일 경우 1 로하고 이하일 경우 0 으로 반환한다.
result_pred <- ifelse(pred_wd2 >= 0.5, 1, 0)
result_pred
table(result_pred)
#모델 평가
table(result_pred, test$label)
(111+50) / nrow(test)
# 정확도 측정 결과값 94%
```

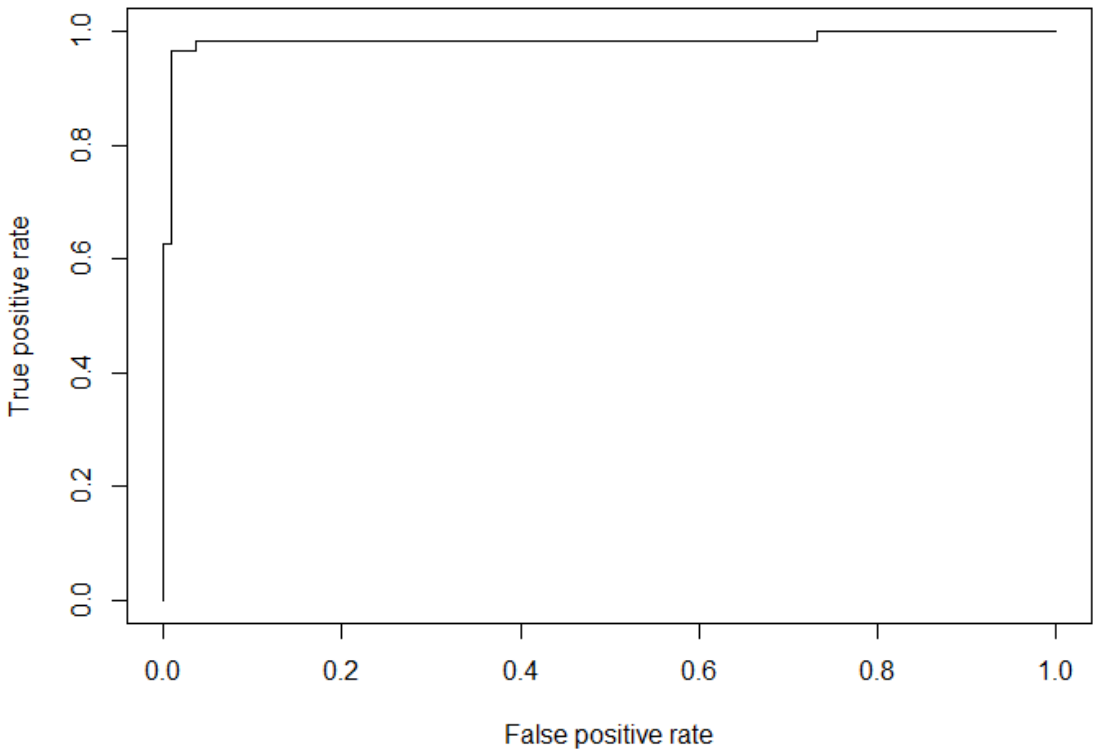
후진제거한 모델로 예측치를 구하고 **pred_wd2** 에 담는다. 다음 분류값을 0.5 이상을 1 로하고 이하는 0 으로 변경하여 **result_pred** 에 담는다.

평가 테이블 실행결과:

	test_lab$label(실제값)	
result_pred (예측값)	0	1
0	**111**	9
1	1	**50**

이에 해당 분류모델의 **Accuracy 정확도는 (111+50)/171 = 0.9415205 약 94%이다**

시각화:



- 로지스틱 회귀분석 결과 분석

해당 로지스틱 회귀 분석 결과는 각각의 변수들이 종양이 악성인지 아닌지를 예측하는 데에 어떤 영향을 미치는지를 보여줍니다.

우선, 절편(intercept)은 -1.59 로 나타나며, 이는 독립변수들이 모두 0 일 때 종양이 악성일 확률을 나타냅니다. radius_mean, area_mean, compactness_mean, points_mean, smoothness_se, concavity_se, symmetry_se, radius_worst, texture_worst, area_worst, concavity_worst, dimension_worst 변수들의 계수를 보면, 각각의 변수들이 종양이 악성일 확률에 미치는 영향을 나타냅니다.

radius_mean, area_mean, compactness_mean, points_mean, concavity_worst, texture_worst, area_worst, radius_worst, dimension_worst 변수들의 계수가 양수이므로 해당 변수의 값이 증가할수록 종양이 악성일 확률이 높아집니다. 반면, smoothness_se 와 concavity_se, points_se 의 계수가 음수이므로 해당 변수의 값이 증가할수록 종양이 악성일 확률이 낮아집니다. symmetry_se 의 계수가 양수이므로 해당 변수의 값이 증가할수록 종양이 악성일 확률이 높아집니다.

또 한,각 변수의 t-valu 보면, radius_mean, area_mean, compactness_mean, points_mean, smoothness_se, concavity_se, symmetry_se, radius_worst, texture_worst, area_worst, concavity_worst, dimension_worst 모두 유의미한 영향을 미치는 것으로 나타납니다.

3. 분석결과 비교

1)xgboost / 로지스틱회귀분석 비교

xgboost 분석 결과 정확도는 92%이고.

로지스틱회귀 분석 결과 정확도는 94% 도출되었다.

두가지 방법 전부 상당히 높은 정확도가 나왔다

데이터의 특성, 모델의 설명력 및 모델의 성능과 속도를 고려하여 더 나은 모델을 선택해야 한다. 일반적으로는, 데이터의 특성이 복잡하고 비선형적인 경우에는 xgboost 와 같은 모델을 사용하는 것이 더 나을 수 있다 하지만, 변수 간의 관계가 단순하고 선형적인 경우에는 로지스틱 회귀분석 모델을 선택하는 것이 더 좋을수 있다.