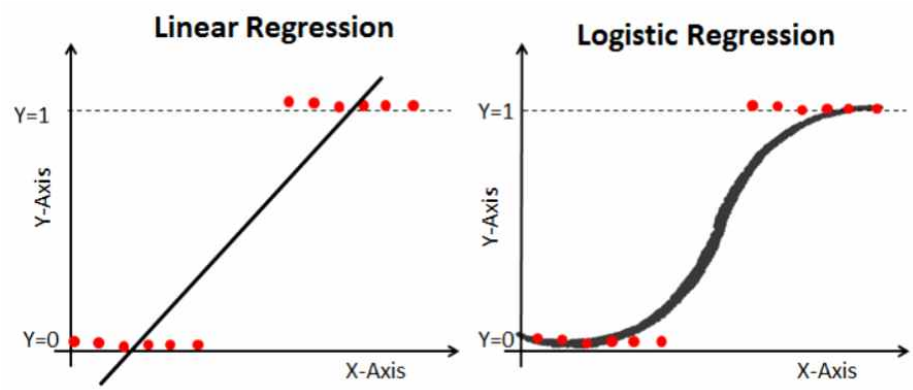
2021-1 YDMS 6주차 과제

〈Subject : Logistic Regression〉

김하은

✓ 이론

로지스틱 회귀분석은 종속변수 Y가 범주형인 경우로 선형회귀의 개념을 확장한 것이다. 특히 범주형 (그 중에서도 이진형) 반응변수를 설명하거나 예측하기 위해 자주 사용되는 회귀분석이다. 이진형 변수의 예시는 대표적으로 [성공/실패],[예/아니오],[구매/비구매],[생존/사망] 등이 있다. 이처럼 이진형 변수는 두 분류로 나눠지며, 이때 이진형 변수는 관측치를 클래스로 분류해주는 변수임을 알 수 있다.



한편 로지스틱 회귀분석을 표현하면 위의 그림과 같다. 로지스틱 회귀분석의 경우 예측'값'을 추정하는 것이 아니라 각 관측치의 '클래스'를 추정하는 것이다. 그러나 범주형 변수가 종속변수인 상황에서 선형 증감으로는 관측치의 클래스를 추정하긴 힘들다. (그림에서 확인할 수 있듯이 선형회귀보다는 비선형회귀에서 더 뚜렷하게 분류됨을 알 수 있다.) 따라서 비선형 증감를 이용하는 것이 로지스틱 회귀분석이다. 이를 위해 사용하는 것이 바로 '로짓(logit)' 변환이다. 이 과정을 이진형 변수에서 연속형 변수로 바꾸어주는 과정이라고 생각하면 이해하기 쉽다.

우선 로짓변환을 설명하기 앞서 몇가지 개념을 설명하려 한다. 우선 q개의 설명변수로 반응변수가 1일 확률은 다음과 같다.

$$p = P(Y=1) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q$$

그러나 이 경우에는 p가 구간 [0,1]에 들어간다는 보장이 안 된다. 이러한 문제를 해결하기 위해 앞서 설명하였듯 비선형 증감을 이용하게 된다. 이러한 문제를 보완하기 위해 Y=1일 확률(p)을 로지스틱 함수를 이용하여 표현하게 된다. 로지스틱 반응함수를 표현하면 다음과 같다.

$$f(x) = \frac{1}{1+e^{-x}}$$

로지스틱 함수는 앞서 언급한 특정 현상을 분류하는 상황에서 매우 적합하다. 한편 로지스틱 함수를 이용하여 p에 관해 다시 표현하면 다음과 같다. 로지스틱 함수를 통해 임의의 변수 $x_1,\cdots,x_q$ 에 대해 p는 항상 구간 [0,1] 사이의 값이 된다.

$$p = P(Y=1|X=x_1,\cdots,x_q) = \frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_qx_q)}}$$

(✓ 과제 : 2. 오즈비에 대해 조사해오기)

이를 이용하여 Odds(오즈)에 대해 알아보자. 오즈는 쉽게 말해 사건 A가 발생할 확률과 그렇지 않을 확률의 비로 정의된다. 즉 클래스 1(Y=1)에 속할 오즈는, 클래스 0에 속하는 확률에 대한 클래스 1에 속하는 확률의 비인 것이다. 오즈에 대한 더 직관적인 이해를 위해 예시를 들어보자. 만약 게임에서 이길 확률이 $\frac{1}{10}$이라면, 게임에서 질 확률은 $\frac{9}{10}$이다. 이때 게임에서 이길 오즈는 바로 $\frac{\frac{1}{10}}{1-\frac{1}{10}}=\frac{1}{9}$ 이다. 이를 수식으로 일반화하면 아래와 같다.

$$Odds = \frac{p_A}{1-p_A}$$

$$Odds(Y=1) = \frac{p}{1-p}$$

한편 역으로 오즈로부터 확률을 구할 수 있다. p를 Odds에 관한 식으로 정리한 것에 로지스틱 함수를 대입하면, 예측변수와 오즈에 관계는 다음과 같이 정리된다.

$$p = \frac{Odds}{1+Odds}$$

$$Odds(Y=1) = e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_qx_q}$$

마지막으로 정리한 Odds식 양변에 자연로그를 취하면 로지스틱 모델의 표준형을 얻을 수 있다. 이때 $\log(Odds)$ 를 로짓 (logit)이라고 부른다. 오즈의 범위는 [0,+∞]인 반면에 오즈에 로그를 취한 로짓의 범위는 [-∞,+∞]이다. 이렇게 회귀식의 변환을 통해 식의 좌변과 우변 모두 [-∞,+∞]에서 정의되기 때문에 앞서 발생한 문제를 해결할 수 있다.

다음으로 Odds Ratio (오즈비)란 Odds의 비율이다. 이를 직관적으로 이해하기 위해 예시를 들어보겠다.

	당첨	당첨X	합계
도심	1	1,999	2,000
비도심	1	7,999	8,000
합계	2	9,998	10,000

도심지역 당첨 오즈(odds)$= \frac{\frac{1}{2000}}{1-\frac{1}{2000}} = \frac{1}{1999}$, 도심지역 당첨 확률$= \frac{1}{2000}$

비 도심지역 당첨 오즈(odds)$= \frac{\frac{1}{8000}}{1-\frac{1}{8000}} = \frac{1}{7999}$, 비 도심지역 당첨 확률$= \frac{1}{8000}$

비 도심지역 대비 도심지역 당첨 오즈비(odds)$= \frac{\frac{1}{1999}}{\frac{1}{7999}} = 4.0015$

오즈와 확률을 비교해보면 값이 거의 같음을 알 수 있다. 마찬가지로 오즈비도 극단적인 경우에 값에 민감하지 않고 안정적이다. 또한 오즈를 오즈로 나누면 비교가 가능해진다. 앞서 제시한 예시에서 비 도심지역 대비 도심지역 당첨 오즈비는 약 4.0015 였다. 즉 비 도심지역에서 1명 당첨될 때 도심지역에서 4명이 당첨됨을 알 수 있다. 이처럼 오즈는 확률과 대조적으로 모델의 결과를 오즈로 보고할 때 어떠한 값에 대해서도 위와 같은 해석이 가능하다. 확률의 경우 특정 예측변수의 한 단위 증가에 따른 확률의 변화가 일정하지 않다. 반면 오즈는 값에 따라 변동하는 것이 크지 않고 안정적이기 때문에 모델의 결과를 오즈로 하는 것이 더 좋다. 한편 오즈비에 관한 해석은 다음과 같다.

① OR=1 인 경우
: 결과와 원인이 유의미하지 않다.
만약 OR의 신뢰구간에 1이 포함된다면 유의미하지 않다고 해석한다.
② OR>1 인 경우
: 원인이 결과에 OR배 영향을 준다.
③ OR<1 인 경우
: 원인이 결과에 1/OR배 영향을 준다.

✓ 과제 (회귀모형)

1. Telecom에 관한 3개의 데이터를 적절히 활용하여 고객 이탈 여부에 대해 어떤 설명변수들이 얼마나 영향을 끼치는지 알아보기 위해 로지스틱 회귀모델을 사용하여 분석하여라.

[데이터 탐색]

01.demographic

: 고객에 대한 인구 통계 정보 - 성별, 연령 범위 및 파트너 및 부양가족이 있는 경우

변수	변수설명
Customer ID (범주형)	고객 ID
gender (범주형)	성별
SeniorCitizen (범주형)	고령 여부
Partner (범주형)	배우자 유무
Dependents (범주형)	부양가족 유무
Churn (범주형)	고객의 이탈 여부, 타겟변수

* 지난 한 달 이내에 퇴사한 고객들 - 이 칼럼은 "Churn"이라고 불립니다.

02.services

: 각 고객이 가입한 서비스 ⇨ 전화, 여러 회선, 인터넷, 온라인 보안, 온라인 백업, 장치 보호, 기술 지원, TV 및 영화 스트리밍

변수	변수설명
Customer ID (범주형)	고객 ID
PhoneServices (범주형)	폰서비스 가입 여부
MultipleLines (범주형)	폰서비스 상품중 MultipleLine 서비스 가입 여부
InternetService (범주형)	인터넷 서비스 가입 여부, 인터넷 서비스 종류
OnlineSecurity (범주형)	인터넷 서비스 상품 중 OnlineSecurity 서비스 가입 여부
OnlineBackup (범주형)	인터넷 서비스 상품 중 OnlineBackup 서비스 가입 여부
DeviceProtection (범주형)	인터넷 서비스 상품 중 DeviceProtection 서비스 가입 여부
TechSupport (범주형)	인터넷 서비스 상품 중 TechSupport 서비스 가입 여부
StreamingTV (범주형)	인터넷 서비스 상품 중 StreamingTV 서비스 가입 여부
StreamingMovies (범주형)	인터넷 서비스 상품 중 StreamingMovies 서비스 가입 여부

03.account

: 고객 계정 정보 ⇨ 고객 활동 기간, 계약, 결제 방법, 페이퍼리스 청구서, 월별 요금 및 총 요금

변수	변수설명
Customer ID (범주형)	고객 ID
tenure (연속형)	데이터 수집 당시까지 해당 고객이 회원으로 머물렀던 총 기간(월)
Contract (범주형)	계약 기간 종류
PaperlessBilling (범주형)	인터넷 청구서 사용 여부
PatmentMethod (범주형)	요금 지불방법에 대한 정보
MonthlyCharges (연속형)	월별 요금
TotalCharges (연속형)	데이터 수집 당시 까지, 총 요금

앞서 설명한 3개의 데이터 집합은 한 통신사에서 고객이 이탈하는 정보를 얻을 수 있는 데이터 셋이다. target 변수를 고객 이탈 여부로 두었으니, 분석의 목적은 고객의 이탈을 예측하는 것이 되겠다. 우선 3개의 데이터를 엮어 분석해야하므로 데이터 셋을 합치기로 하였다. 이 과정에서 데이터 열이 맞지 않음을 알게 되었는데, <01.demographic> 데이터 셋에 존재하는 중복값으로 인해 생긴 문제임을 확인하였다. 따라서 <01.demographic> 속 중복데이터를 삭제해준 뒤 3개의 데이터 셋을 합쳐주었다.

ⓐ 중복데이터 확인

```
> summary(a)
      customerID     gender      SeniorCitizen    Partner     Dependents Churn
 9631-XEYKE:   2   Female:3488   Min.   :0.0000   No :3642   No :4934   No :5175
 0002-ORFBO:   1   Male  :3556   1st Qu.:0.0000   Yes:3402   Yes:2110   Yes:1869
 0003-MKNFE:   1                 Median :0.0000
 0004-TLHLJ:   1                 Mean   :0.1621
 0011-IGKFF:   1                 3rd Qu.:0.0000
 0013-EXCHZ:   1                 Max.   :1.0000
 (Other)   :7037
```

customerID	gender	SeniorCitizen	Partner	Dependen	Churn
9631-XEYKE	Male	0	No	No	No
9631-XEYKE	Male	0	No	No	No

ⓑ 중복데이터 제거

```
> a<-distinct(a)
> summary(a)
      customerID     gender      SeniorCitizen    Partner     Dependents Churn
 0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933   No :5174
 0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110   Yes:1869
 0004-TLHLJ:   1                 Median :0.0000
 0011-IGKFF:   1                 Mean   :0.1621
 0013-EXCHZ:   1                 3rd Qu.:0.0000
 0013-MHZWF:   1                 Max.   :1.0000
 (Other)   :7037
```

ⓒ 데이터 셋 합치기 : (01.demographic) + (02.services) + (03.account)

merge 함수를 이용하여 데이터를 합쳤는데, 기준은 공통 변수인 customerID 를 기준으로 데이터를 병합하였다. str 함수를 이용하여 데이터의 구성을 확인하였는데, 합친 데이터 셋 (t)는 총 7043개의 데이터들과 21개의 변수들로 구성되어있음을 알 수 있었다.

```
> str(t)
'data.frame':   7043 obs. of  21 variables:
 $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 1 2 1 1 ...
 $ SeniorCitizen   : int  0 0 0 1 1 0 1 0 1 0 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 1 2 2 1 2 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 1 1 1 2 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 2 2 1 1 1 1 1 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 1 3 1 1 1 1 1 3 1 3 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 2 2 2 1 2 2 1 1 ...
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 1 3 3 3 3 ...
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 1 3 1 1 3 1 1 3 ...
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 1 3 3 1 1 3 1 1 3 ...
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 3 1 1 1 3 3 3 3 1 3 ...
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 3 1 1 3 3 3 3 1 1 3 ...
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 3 1 1 3 ...
 $ tenure          : int  9 9 4 13 3 9 71 63 7 65 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 2 1 1 1 1 1 3 3 1 3 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 2 2 2 2 2 2 2 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 4 4 3 3 4 2 1 2 3 4 ...
 $ MonthlyCharges  : num  65.6 59.9 73.9 98 83.9 ...
 $ TotalCharges    : num  593 542 281 1238 267 ...
```

다음으로 합친 데이터 셋 (t)의 형태를 summary 함수를 통해 확인하였다. 이때 TotalCharges 에 결측치가 있는 것을 확인하였고, 11개의 값을 삭제해주었다.

ⓓ 결측치 확인

```
> summary(t)
      customerID      gender     SeniorCitizen    Partner    Dependents Churn      PhoneService          MultipleLines     InternetService
 0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933   No :5174   No : 682     No              :3390   DSL        :2421
 0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110   Yes:1869   Yes:6361     No phone service: 681   Fiber optic:3096
 0004-TLHLJ:   1                 Median :0.0000                                                 Yes             :2972   No         :1526
 0011-IGKFF:   1                 Mean   :0.1621
 0013-EXCHZ:   1                 3rd Qu.:0.0000
 0013-MHZWF:   1                 Max.   :1.0000
 (Other)   :7037
       OnlineSecurity             OnlineBackup           DeviceProtection          TechSupport              StreamingTV
 No              :3498   No              :3088   No              :3095   No              :3473   No              :2810
 No internet service:1526   No internet service:1526   No internet service:1526   No internet service:1526   No internet service:1526
 Yes             :2019   Yes             :2429   Yes             :2422   Yes             :2044   Yes             :2707


     StreamingMovies     tenure               Contract    PaperlessBilling                 PaymentMethod  MonthlyCharges    TotalCharges
 No              :2785   Min.   : 0.00   Month-to-month:3875   No :2872         Bank transfer (automatic):1544   Min.   : 18.25   Min.   :   0.0
 No internet service:1526   1st Qu.: 9.00   One year      :1473   Yes:4171         Credit card (automatic) :1522   1st Qu.: 35.50   1st Qu.: 401.4
 Yes             :2732   Median :29.00   Two year      :1695                    Electronic check        :2365   Median : 70.35   Median :1397.5
                         Mean   :32.37                                          Mailed check            :1612   Mean   : 64.76   Mean   :2283.3
                         3rd Qu.:55.00                                                                          3rd Qu.: 89.85   3rd Qu.:3794.7
                         Max.   :72.00                                                                          Max.   :118.75   Max.   :8684.8
                                                                                                                                 NA's   :11
```

ⓔ 결측치 삭제

```
> na.omit(t)->t
> summary(t)
      customerID      gender     SeniorCitizen    Partner    Dependents Churn      PhoneService          MultipleLines     InternetService
 0002-ORFBO:   1   Female:3483   Min.   :0.0000   No :3639   No :4933   No :5163   No : 680     No              :3385   DSL        :2416
 0003-MKNFE:   1   Male  :3549   1st Qu.:0.0000   Yes:3393   Yes:2099   Yes:1869   Yes:6352     No phone service: 679   Fiber optic:3096
 0004-TLHLJ:   1                 Median :0.0000                                                 Yes             :2968   No         :1520
 0011-IGKFF:   1                 Mean   :0.1624
 0013-EXCHZ:   1                 3rd Qu.:0.0000
 0013-MHZWF:   1                 Max.   :1.0000
 (Other)   :7026
       OnlineSecurity             OnlineBackup           DeviceProtection          TechSupport              StreamingTV
 No              :3497   No              :3087   No              :3094   No              :3472   No              :2809
 No internet service:1520   No internet service:1520   No internet service:1520   No internet service:1520   No internet service:1520
 Yes             :2015   Yes             :2425   Yes             :2418   Yes             :2040   Yes             :2703


     StreamingMovies     tenure               Contract    PaperlessBilling                 PaymentMethod  MonthlyCharges    TotalCharges
 No              :2781   Min.   : 1.00   Month-to-month:3875   No :2864         Bank transfer (automatic):1542   Min.   : 18.25   Min.   :   0.0
 No internet service:1520   1st Qu.: 9.00   One year      :1472   Yes:4168         Credit card (automatic) :1521   1st Qu.: 35.59   1st Qu.: 401.4
 Yes             :2731   Median :29.00   Two year      :1685                    Electronic check        :2365   Median : 70.35   Median :1397.5
                         Mean   :32.42                                          Mailed check            :1604   Mean   : 64.80   Mean   :2283.3
                         3rd Qu.:55.00                                                                          3rd Qu.: 89.86   3rd Qu.:3794.7
                         Max.   :72.00                                                                          Max.   :118.75   Max.   :8684.8

> sum(is.na(t))
[1] 0
```

[데이터 모델링]

전처리를 완료한 데이터 셋을 모델링하기 위해 우선 데이터 셋을 train set 과 test set 으로 나누어 주었다. 이때 비율은 7:3 으로 진행하였다.

ⓐ 분할

```
> h<-t[,-c(1)]
> View(h)
> set.seed(1234)
> idx <- sample(x = c("train", "test"),
+               size = nrow(h),
+               replace = TRUE,
+               prob = c(7, 3))
>
> train<-h[idx == "train", ]
> test<-h[idx == "test", ]
> dim(train)
[1] 4954   20
> dim(test)
[1] 2078   20
```

ⓑ train set

다음으로 train set으로 로지스틱 회귀분석을 돌려 전반적인 성능을 확인해보았다. 빨간색 표시선에서 볼 수 있듯 결측치로 나오는 변수가 생겼음을 확인하였다. 공통적으로 3개의 범주로 나누어진 변수에서 발생하였다. 이를 참고하여 "Yes", "No", "No~~service" 으로 나누어진 범주형 변수는 "Yes"와 "No"로만 구성될 수 있도록 수정하였다. 이후 수정된 데이터를 가지고 다시 로지스틱 회귀모델을 돌려보았다. 수정 이후 NA값은 사라진 것을 확인할 수 있다. train의 로지스틱 회귀모델에 사용된 예측변수는 총 23개이다.

수정 전

```
> logit_train<-glm(Churn~.,data=train,family='binomial')
> summary(logit_train)

Call:
glm(formula = Churn ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9319  -0.6788  -0.2917   0.7372   3.3994

Coefficients: (6 not defined because of singularities)
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -7.565e+00  1.970e+02  -0.038 0.969364
genderMale                             1.589e-02  7.709e-02   0.206 0.836747
SeniorCitizen                          2.429e-01  1.022e-01   2.377 0.017477 *
PartnerYes                             4.930e-02  9.312e-02   0.529 0.596498
DependentsYes                         -1.599e-01  1.075e-01  -1.487 0.136904
PhoneServiceYes                        9.937e+00  1.970e+02   0.050 0.959766
MultipleLinesNo phone service          9.414e+00  1.970e+02   0.048 0.961881
MultipleLinesYes                       5.459e-01  2.125e-01   2.568 0.010217 *
InternetServiceFiber optic             2.512e+00  9.575e-01   2.623 0.008708 **
InternetServiceNo                     -2.233e+00  9.656e-01  -2.312 0.020759 *
OnlineSecurityNo internet service             NA         NA      NA       NA
OnlineSecurityYes                     -1.019e-02  2.137e-01  -0.048 0.961959
OnlineBackupNo internet service               NA         NA      NA       NA
OnlineBackupYes                        1.845e-01  2.096e-01   0.880 0.378757
DeviceProtectionNo internet service           NA         NA      NA       NA
DeviceProtectionYes                    2.907e-01  2.098e-01   1.385 0.165983
TechSupportNo internet service                NA         NA      NA       NA
TechSupportYes                        -2.932e-02  2.141e-01  -0.137 0.891039
StreamingTVNo internet service                NA         NA      NA       NA
StreamingTVYes                         8.498e-01  3.898e-01   2.180 0.029271 *
StreamingMoviesNo internet service            NA         NA      NA       NA
StreamingMoviesYes                     8.408e-01  1.912e-01   2.149 0.031612 *
tenure                                -5.763e-02  7.300e-03  -7.894 2.93e-15 ***
ContractOne year                      -6.690e-01  1.255e-01  -5.332 9.73e-08 ***
ContractTwo year                      -1.414e+00  2.097e-01  -6.743 1.55e-11 ***
PaperlessBillingYes                    3.205e-01  8.770e-02   3.654 0.000258 ***
PaymentMethodCredit card (automatic)  -1.675e-01  1.363e-01  -1.229 0.219184
PaymentMethodElectronic check          2.185e-01  1.124e-01   1.944 0.051839 .
PaymentMethodMailed check             -1.787e-01  1.361e-01  -1.313 0.189158
MonthlyCharges                        -6.477e-02  3.801e-02  -1.704 0.088388 .
TotalCharges                           2.878e-04  8.239e-05   3.493 0.000477 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5771.8  on 4953  degrees of freedom
Residual deviance: 4128.3  on 4929  degrees of freedom
AIC: 4178.3

Number of Fisher Scoring iterations: 10
```

수정 후

```
> summary(logit_train)

Call:
glm(formula = Churn ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9318  -0.6787  -0.2921   0.7379   3.4007

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            1.829e+00  9.763e-01   1.874 0.060968 .
genderMale                             1.617e-02  7.709e-02   0.210 0.833819
SeniorCitizen                          2.433e-01  1.022e-01   2.380 0.017327 *
PartnerYes                             4.981e-02  9.311e-02   0.535 0.592712
DependentsYes                         -1.597e-01  1.075e-01  -1.486 0.137206
PhoneServiceYes                        5.083e-01  7.726e-01   0.658 0.510613
MultipleLinesYes                       5.411e-01  2.121e-01   2.551 0.010727 *
InternetServiceFiber optic             2.492e+00  9.558e-01   2.607 0.009122 **
InternetServiceNo                     -2.213e+00  9.640e-01  -2.296 0.021664 *
OnlineSecurityYes                     -1.484e-02  2.133e-01  -0.070 0.944522
OnlineBackupYes                        1.810e-01  2.094e-01   0.864 0.387395
DeviceProtectionYes                    2.862e-01  2.094e-01   1.367 0.171736
TechSupportYes                        -3.294e-02  2.138e-01  -0.154 0.877574
StreamingTVYes                         8.425e-01  3.893e-01   2.164 0.030427 *
StreamingMoviesYes                     8.323e-01  3.904e-01   2.132 0.033016 *
tenure                                -5.773e-02  7.298e-03  -7.911 2.56e-15 ***
ContractOne year                      -6.699e-01  1.255e-01  -5.339 9.37e-08 ***
ContractTwo year                      -1.414e+00  2.097e-01  -6.743 1.55e-11 ***
PaperlessBillingYes                    3.203e-01  8.770e-02   3.652 0.000260 ***
PaymentMethodCredit card (automatic)  -1.674e-01  1.363e-01  -1.228 0.219409
PaymentMethodElectronic check          2.179e-01  1.124e-01   1.939 0.052443 .
PaymentMethodMailed check             -1.791e-01  1.361e-01  -1.315 0.188417
MonthlyCharges                        -6.399e-02  3.794e-02  -1.686 0.091707 .
TotalCharges                           2.889e-04  8.236e-05   3.508 0.000452 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5771.8  on 4953  degrees of freedom
Residual deviance: 4128.5  on 4930  degrees of freedom
AIC: 4176.5

Number of Fisher Scoring iterations: 6
```

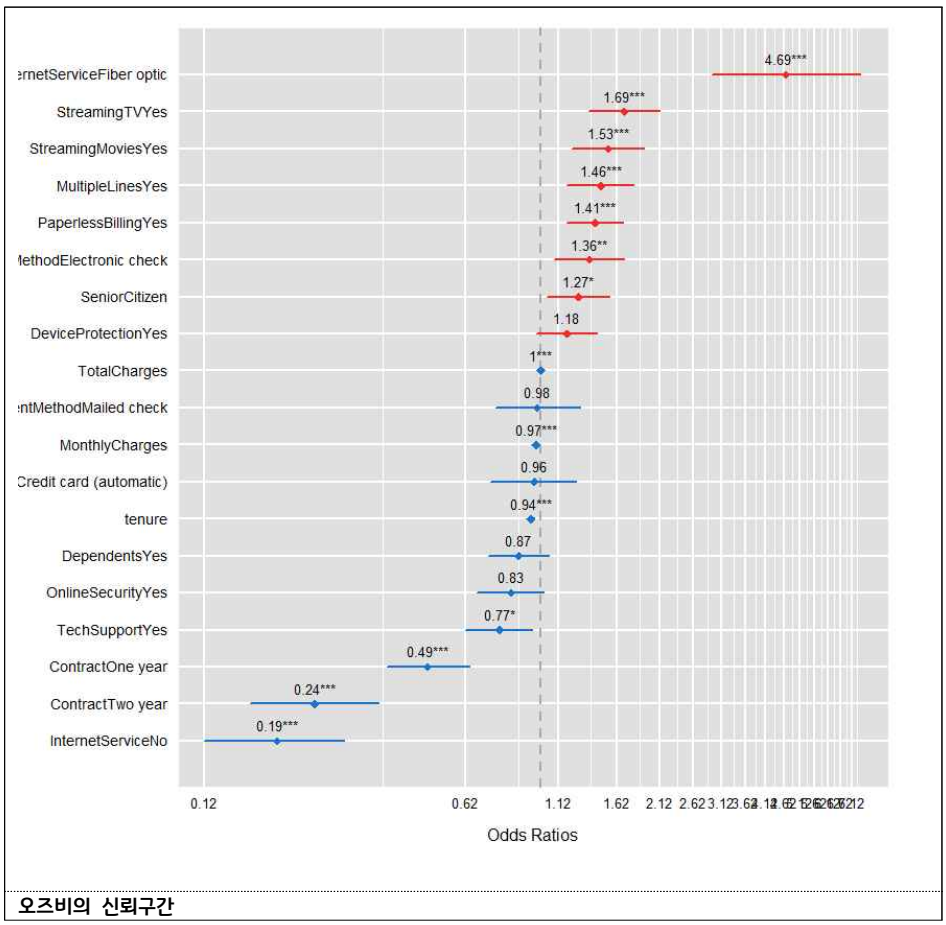
ⓒ train set 의 stepwise

위에서 진행한 모델에서 유용한 변수만 선택하기 위해 단계적 선택방법을 이용하여 모델을 돌려보았다. 그 결과 5개의 변수가 줄어 총 18개의 예측변수가 모델에 사용되었다.

```
> summary(step)

Call:
glm(formula = Churn ~ SeniorCitizen + PhoneService + MultipleLines +
    InternetService + OnlineBackup + DeviceProtection + StreamingTV +
    StreamingMovies + tenure + Contract + PaperlessBilling +
    PaymentMethod + MonthlyCharges + TotalCharges, family = "binomial",
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9345  -0.6800  -0.2940   0.7376   3.4207

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            1.935e+00  4.222e-01   4.582 4.60e-06 ***
SeniorCitizen                          2.727e-01  1.002e-01   2.722 0.006484 **
PhoneServiceYes                        5.865e-01  3.104e-01   1.889 0.058839 .
MultipleLinesYes                       5.657e-01  1.129e-01   5.011 5.43e-07 ***
InternetServiceFiber optic             2.614e+00  3.313e-01   7.890 3.02e-15 ***
InternetServiceNo                     -2.328e+00  4.040e-01  -5.763 8.26e-09 ***
OnlineBackupYes                        2.005e-01  1.129e-01   1.775 0.075880 .
DeviceProtectionYes                    3.087e-01  1.147e-01   2.690 0.007135 **
StreamingTVYes                         8.825e-01  1.629e-01   5.417 6.06e-08 ***
StreamingMoviesYes                     8.782e-01  1.637e-01   5.365 8.12e-08 ***
tenure                                -5.796e-02  7.264e-03  -7.980 1.47e-15 ***
ContractOne year                      -6.818e-01  1.251e-01  -5.451 5.01e-08 ***
ContractTwo year                      -1.437e+00  2.087e-01  -6.885 5.78e-12 ***
PaperlessBillingYes                    3.235e-01  8.754e-02   3.695 0.000220 ***
PaymentMethodCredit card (automatic)  -1.678e-01  1.361e-01  -1.232 0.217770
PaymentMethodElectronic check          2.213e-01  1.122e-01   1.972 0.048653 *
PaymentMethodMailed check             -1.805e-01  1.359e-01  -1.327 0.184351
MonthlyCharges                        -6.852e-02  1.327e-02  -5.162 2.44e-07 ***
TotalCharges                           2.933e-04  8.233e-05   3.563 0.000367 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5771.8  on 4953  degrees of freedom
Residual deviance: 4130.8  on 4935  degrees of freedom
AIC: 4168.8

Number of Fisher Scoring iterations: 6
```

ⓓ-1. 오즈비 확인

```
> extractOR(step)
                                      OR  lcl   ucl      p
(Intercept)                         6.92 3.03 15.83 0.0000
SeniorCitizen                       1.31 1.08  1.60 0.0065
PhoneServiceYes                     1.80 0.98  3.30 0.0588
MultipleLinesYes                    1.76 1.41  2.20 0.0000
InternetServiceFiber optic         13.65 7.13 26.14 0.0000
InternetServiceNo                   0.10 0.04  0.22 0.0000
OnlineBackupYes                     1.22 0.98  1.52 0.0759
DeviceProtectionYes                 1.36 1.09  1.71 0.0071
StreamingTVYes                      2.42 1.76  3.33 0.0000
StreamingMoviesYes                  2.41 1.75  3.32 0.0000
tenure                              0.94 0.93  0.96 0.0000
ContractOne year                    0.51 0.40  0.65 0.0000
ContractTwo year                    0.24 0.16  0.36 0.0000
PaperlessBillingYes                 1.38 1.16  1.64 0.0002
PaymentMethodCredit card (automatic) 0.85 0.65  1.10 0.2178
PaymentMethodElectronic check       1.25 1.00  1.55 0.0487
PaymentMethodMailed check           0.83 0.64  1.09 0.1844
MonthlyCharges                      0.93 0.91  0.96 0.0000
TotalCharges                        1.00 1.00  1.00 0.0004
```

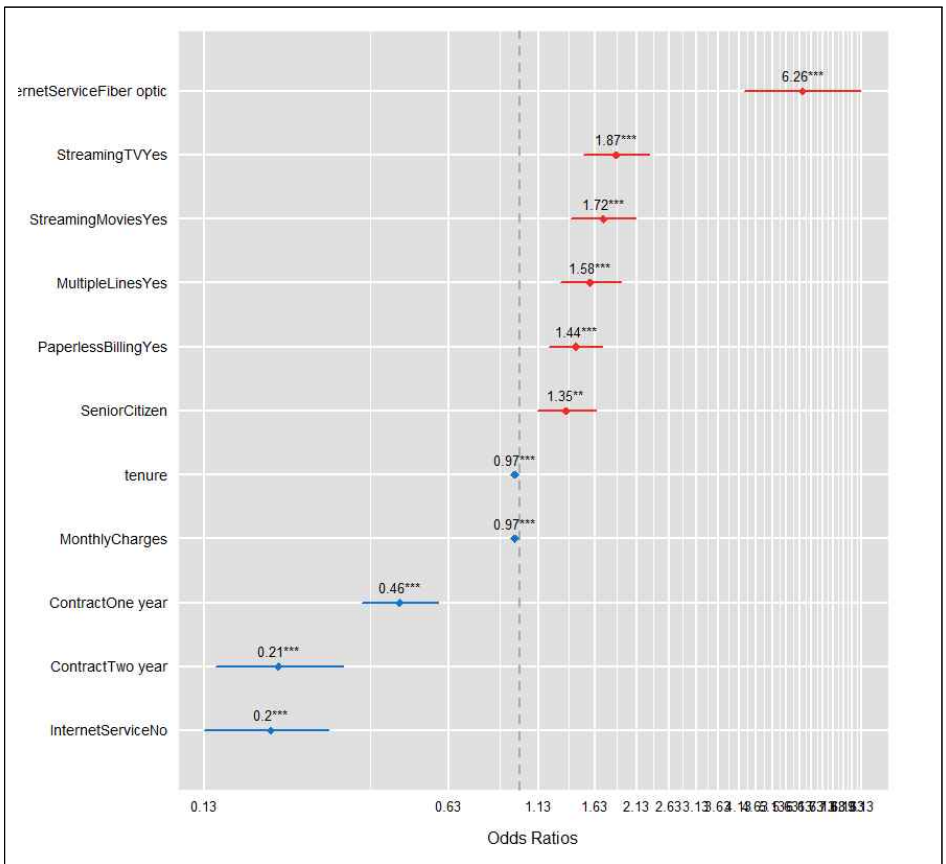


오즈비의 신뢰구간

앞서 이론파트에서 설명한 것과 같이 오즈비 신뢰구간에 1이 포함되어 있으면 유의미하지 않은 변수라고 하였다. 따라서 오즈비 신뢰구간에 1이 들어간 변수는 제거해주었다.

ⓓ-2. '수정된' 모델의 오즈비 확인

```
> extractOR(step.train_1)
                                OR  lcl  ucl      p
(Intercept)                   2.05 1.24 3.38 0.0052
SeniorCitizen                 1.35 1.12 1.64 0.0022
MultipleLinesYes              1.58 1.30 1.93 0.0000
InternetServiceFiber optic    6.26 4.30 9.10 0.0000
InternetServiceNo             0.20 0.13 0.29 0.0000
StreamingTVYes                1.87 1.51 2.32 0.0000
StreamingMoviesYes            1.72 1.39 2.12 0.0000
tenure                        0.97 0.96 0.97 0.0000
ContractOne year              0.46 0.36 0.59 0.0000
ContractTwo year              0.21 0.14 0.32 0.0000
PaperlessBillingYes           1.44 1.21 1.71 0.0000
MonthlyCharges                0.97 0.96 0.98 0.0000
```



'수정된' 모델의 오즈비의 신뢰구간

오즈비의 신뢰구간에 1이 포함되지 않음을 확인할 수 있다. 따라서 최종 선택한 변수는 8개로 다음과 같다. SeniorCitizen, MultipleLines, InternetService, StreamingTV, StreamingMovies, tenure+Contract, PaperlessBilling, MonthlyCharges

ⓔ train set 회귀 모델

선택한 변수를 가지고 회귀모델을 돌렸다. 각 변수의 p-value값이 낮게 나옴을 보고 더 이상 수정하지 않았다.

```
> summary(step.train_1)

Call:
glm(formula = Churn ~ SeniorCitizen + MultipleLines + InternetService +
    StreamingTV + StreamingMovies + tenure + Contract + PaperlessBilling +
    MonthlyCharges, family = "binomial", data = train_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9383  -0.6624  -0.2990   0.7179   3.2141

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.716373   0.256126   2.797  0.00516 **
SeniorCitizen              0.302712   0.098713   3.067  0.00217 **
MultipleLinesYes           0.457355   0.100834   4.536 5.74e-06 ***
InternetServiceFiber optic 1.833389   0.191014   9.598  < 2e-16 ***
InternetServiceNo         -1.627807   0.198603  -8.196 2.48e-16 ***
StreamingTVYes             0.627834   0.109306   5.744 9.26e-09 ***
StreamingMoviesYes         0.540760   0.108297   4.993 5.93e-07 ***
tenure                    -0.035202   0.002556 -13.773  < 2e-16 ***
ContractOne year          -0.775656   0.126332  -6.140 8.26e-10 ***
ContractTwo year          -1.553983   0.208266  -7.462 8.55e-14 ***
PaperlessBillingYes        0.364783   0.088169   4.137 3.51e-05 ***
MonthlyCharges            -0.031892   0.005575  -5.720 1.06e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5743.0  on 4992  degrees of freedom
Residual deviance: 4155.4  on 4981  degrees of freedom
AIC: 4179.4

Number of Fisher Scoring iterations: 6
```

ⓕ train set 회귀 모델 predict

다음으로 회귀모델을 평가하기 위해 predict 함수를 이용하여 Churn을 예측해보았다. 이 과정 이전에 Churn 의 경우 반드시 “Yes” 와 “No”를 1과 0으로 설정해주어야 한다.

```
> t$Churn<-ifelse(t$Churn=="Yes",1,0)
> str(t$Churn)
 num [1:7032] 0 0 0 0 0 0 0 0 0 0 ...
```

```
> pred_1 <- predict(step.train_1, train_1,type="response")
> pred_1
          3           6           7           8           9          11          14          15          16          17          18          19
0.602746295 0.430362636 0.040463702 0.043128530 0.401374778 0.027862411 0.034909185 0.014123737 0.039379136 0.114970019 0.792490021 0.057348954
         20          21          23          24          25          26          27          28          29          30          31          32
0.633712755 0.029574618 0.626944252 0.161095919 0.154302764 0.427628433 0.126105566 0.006581169 0.054031085 0.003979344 0.219542978 0.016886741
         33          34          35          36          37          38          40          41          42          44          45          46
0.068091193 0.009004129 0.080659749 0.018753037 0.309963100 0.249933617 0.501405423 0.436048469 0.088257241 0.073162022 0.201837440 0.016349932
         47          48          49          50          51          52          55          56          57          58          61          62
0.100384800 0.251642010 0.409300116 0.514508482 0.021493748 0.495000840 0.724898398 0.344513169 0.262342222 0.107568838 0.024005262 0.318417988
         63          64          65          67          68          69          70          72          73          74          77          79
0.041410969 0.228833688 0.789858136 0.631977104 0.054767463 0.017531162 0.397487454 0.563372512 0.133753577 0.548224997 0.015236840 0.246237801
         80          81          83          85          87          89          90          92          94          95          96         100
0.186942258 0.028609345 0.523348111 0.671505910 0.685456732 0.556200141 0.491773211 0.560022939 0.563018344 0.194373434 0.038700162 0.544036590
        101         102         103         104         105         106         108         109         110         114         116         117
0.160665384 0.309910645 0.030257507 0.079910648 0.786011997 0.315730682 0.055436664 0.291619584 0.142321495 0.052611429 0.060130519 0.051336100
        118         119         120         121         122         124         125         127         128         129         130         131
0.012946966 0.142766494 0.459707018 0.124273807 0.467430984 0.303641648 0.090623415 0.628702308 0.656328982 0.090250947 0.576545152 0.302718287
        133         135         136         138         139         141         142         145         146         148         149         150
0.032620221 0.573568181 0.198724977 0.042062809 0.018461530 0.114889783 0.066398628 0.328147622 0.009883675 0.009004199 0.171517854 0.476398052
        151         152         153         154         155         158         159         160         161         162         164         165
0.046714517 0.031021333 0.029019523 0.042172006 0.017464293 0.467177164 0.525448854 0.768130420 0.403237481 0.032416385 0.539773310 0.485903743
        166         168         170         171         173         174         176         178         179         180         181         182
0.060572463 0.161527385 0.005809253 0.440707755 0.026480509 0.146944273 0.007001347 0.165841429 0.079675317 0.125328768 0.811717041 0.376011701
        183         185         187         188         189         190         191         192         193         194         195         196
0.129349468 0.020362091 0.552522396 0.070604472 0.208072575 0.225931215 0.698951927 0.054275970 0.102553140 0.202593980 0.316709419 0.023437318
        198         199         200         201         203         204         206         207         208         209         210         211
0.330236700 0.583187870 0.222063430 0.499714881 0.011338229 0.433232672 0.522254000 0.309954623 0.376976783 0.069730673 0.321665892 0.685971787
        212         213         215         219         223         224         225         226         229         231         232         233
0.140206665 0.238298474 0.701680101 0.036272896 0.014590707 0.659078801 0.108531494 0.111637858 0.535058715 0.082498519 0.080844587 0.016063331
```

이하 생략

ⓖ cutoff 값 설정

지금 하고 있는 회귀모델은 로지스틱 회귀분석 모델이기 때문에 cutoff (컷오프) 값을 지정하여 각 예측값들을 0 또는 1로 클래스를 나누어 주어야 한다. 따라서 cutoff=0.5 로 설정하고 분석을 진행하였다.

```
> predict_1 <- ifelse(pred_1>0.5, 1, 0)
> predict_1
  3   6   7   8   9  11  14  15  16  17  18  19  20  21  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38
  1   0   0   0   0   0   0   0   0   0   1   0   1   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 40  41  42  44  45  46  47  48  49  50  51  52  55  56  57  58  61  62  63  64  65  67  68  69  70  72  73  74  77  79
  1   0   0   0   0   0   0   0   0   1   0   0   1   0   0   0   0   0   0   0   1   1   0   0   0   1   0   1   0   0
 80  81  83  85  87  89  90  92  94  95  96 100 101 102 103 104 105 106 108 109 110 114 116 117 118 119 120 121 122 124
  0   0   1   1   1   1   0   1   1   0   0   1   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
125 127 128 129 130 131 133 135 136 138 139 141 142 145 146 148 149 150 151 152 153 154 155 158 159 160 161 162 164 165
  0   1   1   0   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   1   0   0   1   0
166 168 170 171 173 174 176 178 179 180 181 182 183 185 187 188 189 190 191 192 193 194 195 196 198 199 200 201 203 204
  0   0   0   0   0   0   0   0   0   0   1   0   0   0   1   0   0   0   1   0   0   0   0   0   0   1   0   0   0   0
206 207 208 209 210 211 212 213 215 219 223 224 225 226 229 231 232 233 235 236 237 238 241 242 243 245 246 247 249 252
  1   0   0   0   0   1   0   0   1   0   0   1   0   0   1   0   0   0   0   0   0   1   0   0   0   0   0   1   1   1
253 255 256 257 258 260 261 262 263 264 265 266 267 269 270 272 274 277 278 280 281 282 283 284 285 286 287 288 289 290
  1   0   0   0   0   0   0   1   1   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   1   0   0   0   0   0
291 292 293 294 295 296 297 298 299 301 302 303 304 306 308 309 310 311 312 313 315 316 317 318 319 320 321 322 323 325
  0   0   0   0   1   0   0   1   0   0   0   0   0   1   0   1   0   0   0   0   0   0   0   0   0   0   0   1   0   0
327 329 331 332 333 334 335 337 338 339 340 341 342 343 345 346 347 349 351 352 354 355 356 357 359 360 361 365 366 367
  0   0   0   0   1   0   1   1   0   1   0   1   0   1   1   0   0   0   0   0   0   1   0   0   0   1   0   0   0   0
369 374 375 377 378 380 382 384 385 386 388 390 391 393 394 395 396 397 399 400 401 402 403 404 405 406 407 408 409 410
  1   0   0   0   0   1   0   0   0   0   0   1   0   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0
413 415 417 418 419 420 421 422 423 425 426 427 428 429 431 433 434 435 436 437 440 441 442 443 444 445 447 448 449 450
  1   1   0   0   0   1   0   0   0   0   0   0   0   0   0   1   0   1   0   1   1   0   1   1   1   0   0   0   0   0
451 452 453 454 455 456 457 459 463 464 465 466 467 469 472 473 474 475 476 477 478 479 480 481 482 483 484 485 487 488
  0   1   0   0   0   0   0   0   0   0   1   0   1   1   0   1   0   0   0   0   0   1   0   0   0   1   0   0   0   1
489 493 497 498 499 500 503 505 507 508 510 511 512 513 516 518 521 522 524 527 528 529 530 531 532 537 538 539 541 542
  0   1   0   0   1   0   0   0   0   0   0   0   0   0   1   0   0   0   0   1   0   0   1   0   1   0   0   1   0   0
```

이하 생략

위와 같이 값들이 0 또는 1로 올바르게 바뀌었음을 확인할 수 있다.

[모델 성능 평가]

ⓐ train set 평가

```
> confusionMatrix(predict_1, train_1$Churn)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3317  593
         1  368  715

               Accuracy : 0.8075
                 95% CI : (0.7963, 0.8184)
    No Information Rate : 0.738
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.473

 Mcnemar's Test P-Value : 4.981e-13

            Sensitivity : 0.9001
            Specificity : 0.5466
         Pos Pred Value : 0.8483
         Neg Pred Value : 0.6602
             Prevalence : 0.7380
         Detection Rate : 0.6643
   Detection Prevalence : 0.7831
      Balanced Accuracy : 0.7234

       'Positive' Class : 0
```

train set 에서 정확도는 0.8075 정도이다. 좀더 자세히 보면, 민감도와 특이도도 확인할 수 있다. 이 데이터의 경우 0을 0으로 맞추는 경우를 민감도로 보았다. 하지만 실제로 우리가 생각하는 민감도는 1을 1로 맞추는 경우, 즉 탈퇴한 사람을 잘 찾아내는 것이 중요하므로 다음 분석 결과는 민감도와 특이도를 바꾸어 해석할 생각이다. 이에 맞게 보면 민감도는 0.5466 이고, 특이도는 0.9001이다.

ⓑ test set 으로 모델 돌린 뒤 cutoff (0.5) 설정

```
> pred_test_2 <- predict(step.train_1, test_2,type="response")
> pred_test_2
        1         2         4         5        10        12        13        22        39        43        53        54
0.186093201 0.374569949 0.690575347 0.743406585 0.017642278 0.036507842 0.631715644 0.543958343 0.028325677 0.024094880 0.514508422 0.450089766
       59        60        66        71        75        76        78        82        84        86        88        91
0.149453985 0.150501349 0.565405101 0.177275125 0.070201895 0.683577885 0.580020064 0.164294166 0.074193264 0.066219996 0.161095919 0.108194379
       93        97        98        99       107       111       112       113       115       123       126       132
0.654814120 0.027412041 0.636701039 0.004888883 0.165029458 0.698854354 0.658542229 0.788193833 0.091133965 0.018405389 0.033098102 0.376134193
      134       137       140       143       144       147       156       157       163       167       169       172
0.475182966 0.005846207 0.612716272 0.685426029 0.737951019 0.214318598 0.471231175 0.125165307 0.003662966 0.039418355 0.056467887 0.423712788
      175       177       184       186       197       202       205       214       216       217       218       220
0.379513039 0.506751059 0.025549055 0.020501494 0.134550371 0.718752310 0.029243057 0.655593366 0.121211204 0.535666552 0.726034486 0.022083597
      221       222       227       228       230       234       239       240       244       248       250       251
0.008046859 0.427290165 0.319208912 0.019837321 0.316498563 0.010984899 0.359432014 0.230779398 0.285322671 0.457792913 0.176326926 0.029657667
      254       259       268       271       273       275       276       279       300       305       307       314
0.017565385 0.092988626 0.557425021 0.051453473 0.004857364 0.709618579 0.255325333 0.423088609 0.110982605 0.180568315 0.060859955 0.211815593
      324       326       328       330       336       344       348       350       353       358       362       363
0.315545891 0.088035211 0.264649521 0.032186159 0.706472951 0.008592837 0.653068691 0.149795983 0.233584200 0.536379165 0.008171388 0.349435537
      364       368       370       371       372       373       376       379       381       383       387       389
0.271567533 0.027048460 0.393989540 0.118659298 0.192326642 0.286603528 0.023090113 0.012624982 0.540643217 0.063496936 0.416792363 0.164411850
      392       398       411       412       414       416       424       430       432       438       439       446
0.415681118 0.260204806 0.007713840 0.631977455 0.005612536 0.016694671 0.145685151 0.015820483 0.119852940 0.162179634 0.003436986 0.762615484
      458       460       461       462       468       470       471       486       490       491       492       494
0.226590305 0.526724508 0.482730161 0.018244274 0.514702834 0.628483016 0.025089594 0.382619553 0.299421139 0.653570694 0.758614037 0.088165520
      495       496       501       502       504       506       509       514       515       517       519       520
0.131173668 0.100205055 0.462366101 0.021808089 0.253291775 0.027798293 0.563317236 0.552767329 0.066059999 0.087966419 0.037656119 0.722298572
```

이하 생략

```
> pred_test_2
   1    2    4    5   10   12   13   22   39   43   53   54   59   60   66   71   75   76   78   82   84   86   88   91   93   97   98   99  107  111
   0    0    1    1    0    0    1    1    0    0    1    0    0    0    1    0    0    1    1    0    0    0    0    0    1    0    1    0    0    1
 112  113  115  123  126  132  134  137  140  143  144  147  156  157  163  167  169  172  175  177  184  186  197  202  205  214  216  217  218  220
   1    1    0    0    0    0    0    0    1    1    1    0    0    0    0    0    0    0    0    1    0    0    0    1    0    1    0    1    1    0
 221  222  227  228  230  234  239  240  244  248  250  251  254  259  268  271  273  275  276  279  300  305  307  314  324  326  328  330  336  344
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    1    0    0    0    0    0    0    0    0    0    1    1    0
 348  350  353  358  362  363  364  368  370  371  372  373  376  379  381  383  387  389  392  398  411  412  414  416  424  430  432  438  439  446
   1    0    0    1    0    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0    0    1    0    0    0    0    0    0    0    1
 458  460  461  462  468  470  471  486  490  491  492  494  495  496  501  502  504  506  509  514  515  517  519  520  523  525  526  533  534  535
   0    1    0    0    1    1    0    0    0    1    1    0    0    0    0    0    0    0    1    1    0    0    0    1    0    0    1    0    0    0
 536  540  545  554  562  570  572  576  581  583  584  587  595  597  598  600  601  606  609  613  615  616  618  632  633  634  636  639  645  647
   1    0    0    0    0    0    0    0    0    0    0    0    1    0    1    1    0    1    0    1    0    0    0    0    1    1    0    0    0    0
 648  649  650  653  656  657  658  659  669  675  676  685  691  694  698  699  704  705  707  708  710  711  713  714  715  725  735  736  738  739
   0    1    1    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0    1    0    0    0
 741  751  753  756  757  773  775  780  782  787  790  791  792  799  802  805  819  824  843  853  862  863  864  867  869  870  880  889  893  896
   0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    1    0    0    1    0    1    1    0    1    0    1    0    0    0
 899  916  917  920  923  927  930  936  940  950  952  957  959  961  962  964  969  970  974  977  988  991  994  997 1001 1007 1010 1011 1018 1020
   0    1    0    0    0    0    0    0    0    1    1    1    0    0    0    0    0    0    0    0    0    0    0    0    1    0    1    0    0    0
1029 1039 1045 1046 1049 1058 1059 1060 1064 1065 1066 1070 1073 1074 1080 1082 1089 1090 1091 1092 1095 1098 1099 1100 1103 1105 1107 1109 1111 1115
   0    0    0    1    0    0    0    0    0    0    0    0    0    1    1    0    1    0    0    0    0    0    1    0    1    0    1    0    0    1
1117 1119 1120 1122 1125 1129 1131 1133 1136 1138 1139 1143 1144 1145 1146 1148 1149 1150 1151 1161 1162 1164 1170 1171 1174 1180 1184 1189 1191 1202
   1    0    0    1    1    0    1    0    1    1    0    0    0    0    0    0    1    0    0    1    0    0    1    0    0    0    1    1    1    0
1207 1209 1215 1218 1222 1225 1226 1230 1236 1243 1245 1249 1253 1254 1256 1257 1262 1265 1266 1268 1277 1278 1281 1282 1288 1297 1298 1303 1305 1308
   0    0    0    0    0    0    0    0    0    0    0    1    0    0    1    0    0    0    1    0    0    0    0    0    0    0    0    0    0    0
```

이하 생략

ⓒ test set 평가

```
> confusionMatrix(pred_test_2, test_2$Churn)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1325  266
         1  153  295

               Accuracy : 0.7945
                 95% CI : (0.7763, 0.8119)
    No Information Rate : 0.7249
    P-Value [Acc > NIR] : 2.490e-13

                  Kappa : 0.4505

 Mcnemar's Test P-Value : 4.461e-08

            Sensitivity : 0.8965
            Specificity : 0.5258
         Pos Pred Value : 0.8328
         Neg Pred Value : 0.6585
             Prevalence : 0.7249
         Detection Rate : 0.6498
   Detection Prevalence : 0.7803
      Balanced Accuracy : 0.7112

       'Positive' Class : 0
```

test set 에서의 정확도는 0.7945 이다. train set 에서보다 정확도가 다소 떨어지긴 했지만 큰 차이는 보이지 않음을 보아 train 에서의 과적합은 일어나지 않았다고 판단하였다. 한편 test 에서의 민감도는 0.5258 이며, 특이도는 0.8965 이다.

ⓓ test set에서의 cutoff 조절

한편 컷오프의 값은 어느 정도 분석자의 주관에 따라 설정되는 값이다. 컷오프가 어느 값에서 설정되느냐에 따라 분석의 결과가 달라질 수 있기 때문에 컷오프 설정은 중요하다. 이러한 생각을 하게 되어 test에서의 컷오프의 값을 변경하여 모델을 돌려보았다.

```
> pred_test_3 <- predict(step.train_1, test_2,type="response")
> pred_test_3
          1           2           4           5          10          12          13          22          39          43          53          54
0.186093201 0.374569949 0.690575347 0.743406585 0.017642278 0.036507842 0.631715644 0.543958343 0.028325677 0.024094880 0.514508422 0.450089766
         59          60          66          71          75          76          78          82          84          86          88          91
0.149453985 0.150501349 0.565405101 0.177275125 0.070201895 0.683577885 0.580020064 0.164294166 0.074193264 0.066219996 0.161095919 0.108194379
         93          97          98          99         107         111         112         113         115         123         126         132
0.654814120 0.027412041 0.636701039 0.004888883 0.165029458 0.698854354 0.658542229 0.788193833 0.091133965 0.018405389 0.033098102 0.376134193
        134         137         140         143         144         147         156         157         163         167         169         172
0.475182966 0.005846207 0.612716272 0.685426029 0.737951019 0.214318598 0.471231175 0.125165307 0.003662966 0.039418355 0.056467887 0.423712788
        175         177         184         186         197         202         205         214         216         217         218         220
0.379513039 0.506751059 0.025549055 0.020501494 0.134550371 0.718752310 0.029243057 0.655593366 0.121211204 0.535666552 0.726034486 0.022083597
        221         222         227         228         230         234         239         240         244         248         250         251
0.008046859 0.427290165 0.319208912 0.019837321 0.316498563 0.010984899 0.359432014 0.230779398 0.285322671 0.457792913 0.176326926 0.029657667
        254         259         268         271         273         275         276         279         300         305         307         314
0.017565385 0.092988626 0.557425021 0.051453473 0.004857364 0.709618579 0.255325333 0.423088609 0.110982605 0.180568315 0.060859955 0.211815593
        324         326         328         330         336         344         348         350         353         358         362         363
0.315545891 0.088035211 0.264649521 0.032186159 0.706472951 0.008592837 0.653068691 0.149795983 0.233584200 0.536379165 0.008171388 0.349435537
        364         368         370         371         372         373         376         379         381         383         387         389
0.271567533 0.027048460 0.393989540 0.118659298 0.192326642 0.286603528 0.023090113 0.012624982 0.540643217 0.063496936 0.416792363 0.164411850
        392         398         411         412         414         416         424         430         432         438         439         446
0.415681118 0.260204806 0.007713840 0.631977455 0.005612536 0.016694671 0.145685151 0.015820483 0.119852940 0.162179634 0.003436986 0.762615484
        458         460         461         462         468         470         471         486         490         491         492         494
0.226590305 0.526724508 0.482730161 0.018244274 0.514702834 0.628483016 0.025089594 0.382619553 0.299421139 0.653570694 0.758614037 0.088165520
        495         496         501         502         504         506         509         514         515         517         519         520
0.131173668 0.100205055 0.462366101 0.021808089 0.253291775 0.027798293 0.563317236 0.552767329 0.066059999 0.087966419 0.037656119 0.722298572
```

이하 생략

***컷오프의 값을 0.5에서 0.4로 조정해보았다.**

```
> pred_test_3<-ifelse(pred_test_3>0.4, 1, 0)
> pred_test_3
  1   2   4   5  10  12  13  22  39  43  53  54  59  60  66  71  75  76  78  82  84  86  88  91  93  97  98  99 107 111
  0   0   1   1   0   0   1   1   0   0   1   1   0   0   1   0   0   1   1   0   0   0   0   0   1   0   1   0   0   1
112 113 115 123 126 132 134 137 140 143 144 147 156 157 163 167 169 172 175 177 184 186 197 202 205 214 216 217 218 220
  1   1   0   0   0   0   1   0   1   1   1   0   1   0   0   0   0   1   0   1   0   0   0   1   0   1   0   1   1   0
221 222 227 228 230 234 239 240 244 248 250 251 254 259 268 271 273 275 276 279 300 305 307 314 324 326 328 330 336 344
  0   1   0   0   0   0   0   0   0   1   0   0   0   0   1   0   0   1   0   1   0   0   0   0   0   0   0   0   1   0
348 350 353 358 362 363 364 368 370 371 372 373 376 379 381 383 387 389 392 398 411 412 414 416 424 430 432 438 439 446
  1   0   0   1   0   0   0   0   0   0   0   0   0   0   1   0   1   0   1   0   0   1   0   0   0   0   0   0   0   1
458 460 461 462 468 470 471 486 490 491 492 494 495 496 501 502 504 506 509 514 515 517 519 520 523 525 526 533 534 535
  0   1   1   0   1   1   0   0   0   1   1   0   0   0   1   0   0   0   1   1   0   0   0   1   0   0   1   0   0   0
536 540 545 554 562 570 572 576 581 583 584 587 595 597 598 600 601 606 609 613 615 616 618 632 633 634 636 639 645 647
  1   0   0   0   0   0   0   0   0   0   0   0   1   0   1   1   0   1   0   1   1   0   0   1   1   1   0   0   0   1
648 649 650 653 656 657 658 659 669 675 676 685 691 694 698 699 704 705 707 708 710 711 713 714 715 725 735 736 738 739
  0   1   1   0   0   0   0   0   1   0   1   0   0   0   1   0   0   0   0   0   0   1   0   0   0   0   1   0   0   0
741 751 753 756 757 773 775 780 782 787 790 791 792 799 802 805 819 824 843 853 862 863 864 867 869 870 880 889 893 896
  0   1   0   0   0   0   0   0   0   1   0   0   0   0   1   0   1   0   0   1   1   1   1   0   1   0   1   0   0   1
899 916 917 920 923 927 930 936 940 950 952 957 959 961 962 964 969 970 974 977 988 991 994 997 1001 1007 1010 1011 1018 1020
  0   1   0   1   0   0   0   0   0   1   1   1   1   0   0   1   0   0   0   0   0   1   0   0   1   0   1   0   0   0
1029 1039 1045 1046 1049 1058 1059 1060 1064 1065 1066 1070 1073 1074 1080 1082 1089 1090 1091 1092 1095 1098 1099 1100 1103 1105 1107 1109 1111 1115
   0    0    0    1    0    0    0    0    1    0    0    1    0    1    1    0    1    0    0    0    0    0    1    0    1    0    1    0    0    1
1117 1119 1120 1122 1125 1129 1131 1133 1136 1138 1139 1143 1144 1145 1146 1148 1149 1150 1151 1161 1162 1164 1170 1171 1174 1180 1184 1189 1191 1202
   1    0    0    1    1    0    1    0    1    1    0    0    0    1    0    0    1    0    0    1    0    0    1    0    0    0    1    1    1    0
```

이하 생략

```
> confusionMatrix(pred_test_3, test_2$Churn)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1221  195
         1  257  366

               Accuracy : 0.7783
                 95% CI : (0.7597, 0.7962)
    No Information Rate : 0.7249
    P-Value [Acc > NIR] : 1.93e-08

                  Kappa : 0.4627

 Mcnemar's Test P-Value : 0.004115

            Sensitivity : 0.8261
            Specificity : 0.6524
         Pos Pred Value : 0.8623
         Neg Pred Value : 0.5875
             Prevalence : 0.7249
         Detection Rate : 0.5988
   Detection Prevalence : 0.6945
      Balanced Accuracy : 0.7393

       'Positive' Class : 0
```

위의 분석 결과를 보면 정확도는 다소 0.7783으로 떨어졌지만, 민감도가 0.6524로 증가함을 볼 수 있다. 분석의 목적이 고객의 이탈 여부를 예측하는 것이므로 상당히 의미있는 결과라고 생각된다. 따라서 컷오프의 설정이 분석에서 중요하다고 생각된다.

Discussion

로지스틱 회귀분석에서 다중공선성을 보아야하는가에 대해 이야기를 나눠보고 싶습니다.