과제 #5 분류(classification)

성원호

6/2/2021

통신회사 고객이탈 예측

Logistic regression과 Naive Bayes Classifier 사용

- 종속변수는 Churn(지난 달 이탈 여부)와 독립변수는 고객이 이용중인 서비스(ex.phone, multiple lines), 고객계정 관련 정보(ex. contract, payment method), 고객의 인구 통계학적 정보(ex.gender, age)를 사용하였다.

데이터 수집

```
# install.packages("Epi")
# install.packages("klaR")
# install.packages("e1071")
library(Epi)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(klaR)
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(e1071)
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(gmodels)
# https://www.kaggle.com/blastchar/telco-customer-churn 로부터 데이터를 가져옴
telco <- read.csv("WA_Fn-UseC_-Telco-Customer-Churn.csv"); head(telco)
```

```
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1 7590-VHVEG Female             0     Yes         No      1           No
## 2 5575-GNVDE   Male             0      No         No     34          Yes
## 3 3668-QPYBK   Male             0      No         No      2          Yes
## 4 7795-CFOCW   Male             0      No         No     45           No
## 5 9237-HQITU Female             0      No         No      2          Yes
## 6 9305-CDSKC Female             0      No         No      8          Yes
##      MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection
## 1 No phone service             DSL             No          Yes               No
## 2               No             DSL            Yes           No              Yes
## 3               No             DSL            Yes          Yes               No
## 4 No phone service             DSL            Yes           No              Yes
## 5               No     Fiber optic             No           No               No
## 6              Yes     Fiber optic             No           No              Yes
##   TechSupport StreamingTV StreamingMovies       Contract PaperlessBilling
## 1          No          No              No Month-to-month              Yes
## 2          No          No              No       One year               No
## 3          No          No              No Month-to-month              Yes
## 4         Yes          No              No       One year               No
## 5          No          No              No Month-to-month              Yes
## 6          No         Yes             Yes Month-to-month              Yes
##               PaymentMethod MonthlyCharges TotalCharges Churn
## 1          Electronic check          29.85        29.85    No
## 2              Mailed check          56.95      1889.50    No
## 3              Mailed check          53.85       108.15   Yes
## 4 Bank transfer (automatic)          42.30      1840.75    No
## 5          Electronic check          70.70       151.65   Yes
## 6          Electronic check          99.65       820.50   Yes
```

데이터 정제

```
# 여자는 1 남자는 0
telco$gender <- as.factor(ifelse(telco$gender=="Female", '1', '0'))

# Yes는 1 No는 0
telco$Partner <- as.factor(ifelse(telco$Partner=="Yes", '1', '0'))

telco$Dependents <- as.factor(ifelse(telco$Dependents=="Yes", '1', '0'))

telco$PhoneService <- as.factor(ifelse(telco$PhoneService=="Yes", '1', '0'))

telco$OnlineSecurity <- as.factor(ifelse(telco$OnlineSecurity=="Yes", '1', '0'))

telco$OnlineBackup <- as.factor(ifelse(telco$OnlineBackup=="Yes", '1', '0'))

telco$DeviceProtection <- as.factor(ifelse(telco$DeviceProtection=="Yes", '1', '0'))

telco$TechSupport <- as.factor(ifelse(telco$TechSupport=="Yes", '1', '0'))

telco$StreamingTV <- as.factor(ifelse(telco$StreamingTV=="Yes", '1', '0'))

telco$StreamingMovies <- as.factor(ifelse(telco$StreamingMovies=="Yes", '1', '0'))

telco$PaperlessBilling <- as.factor(ifelse(telco$PaperlessBilling=="Yes", '1', '0'))

telco$Churn <- as.factor(ifelse(telco$Churn=="Yes", '1', '0'))

# 데이터 프레임 내 수치형 자료들은 따로 추출하여 numeric 형태로 변환
num_columns <- c("tenure", "MonthlyCharges", "TotalCharges")
telco[num_columns] <- sapply(telco[num_columns], as.numeric)

# scale 함수를 이용해 표준화
telco_int <- telco[,c("tenure", "MonthlyCharges", "TotalCharges")]
telco_int <- data.frame(scale(telco_int))

# Customer No와 수치형 자료들 제거
telco_abs <- telco[,-c(1,6,19,20)]


# 이진 표기로 변환이 안되는 변수들은 더미 변수화 진행
telco_dummy_1 <- transform(telco_abs,
                           MultipleLines_no = ifelse(MultipleLines == "No", 1, 0)) # 코드
를 돌리는데 자꾸 논리값이 하나만 나오는 오류가 발생하여 소거하고 하나로만 진행했습니다

telco_dummy_2 <- transform(telco_dummy_1,
                           Pay_Mailed = ifelse(PaymentMethod == "Mailed check", 1, 0),
                           Pay_Credit = ifelse(PaymentMethod == "Credit card (automati
c)", 1, 0))

telco_dummy_3 <- transform(telco_dummy_2,
                           Contract_Month = ifelse(Contract == "Month-to-month", 1, 0),
                           Contract_One_year = ifelse(Contract == "One year", 1, 0))

telco_dummy_4 <- transform(telco_dummy_3,
```

```
                            Service_DSL = ifelse(InternetService == "DSL", 1, 0),
                            Service_Fiber = ifelse(InternetService == "Fiber optic", 1, 0
))

# 더미 변수화 이후 원 데이터 컬럼들은 제거
rev_telco <- telco_dummy_4[,-c(6, 7, 14, 16)]

# 표준화 했던 수치형 자료들과 결합
final_telco <- cbind(rev_telco, telco_int)

# 정제한 데이터들을 범주형으로 변환(논리 값이 1과 0만 존재하도록 만들었기 때문에)
final_telco$MultipleLines_no <- as.factor(final_telco$MultipleLines_no)
final_telco$Pay_Mailed <- as.factor(final_telco$Pay_Mailed)
final_telco$Pay_Credit <- as.factor(final_telco$Pay_Credit)
final_telco$Contract_Month <- as.factor(final_telco$Contract_Month)
final_telco$Contract_One_year <- as.factor(final_telco$Contract_One_year)
final_telco$Service_DSL <- as.factor(final_telco$Service_DSL)
final_telco$Service_Fiber <- as.factor(final_telco$Service_Fiber)
final_telco$SeniorCitizen <- as.factor(final_telco$SeniorCitizen)

head(final_telco); str(final_telco)
```

```
##   gender SeniorCitizen Partner Dependents PhoneService OnlineSecurity
## 1      1             0       1          0            0              0
## 2      0             0       0          0            1              1
## 3      0             0       0          0            1              1
## 4      0             0       0          0            0              1
## 5      1             0       0          0            1              0
## 6      1             0       0          0            1              0
##   OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies
## 1            1                0           0           0               0
## 2            0                1           0           0               0
## 3            1                0           0           0               0
## 4            0                1           1           0               0
## 5            0                0           0           0               0
## 6            0                1           0           1               1
##   PaperlessBilling Churn MultipleLines_no Pay_Mailed Pay_Credit Contract_Month
## 1                1     0                0          0          0              1
## 2                0     0                1          1          0              0
## 3                1     1                1          1          0              1
## 4                0     0                0          0          0              0
## 5                1     1                1          0          0              1
## 6                1     1                0          0          0              1
##   Contract_One_year Service_DSL Service_Fiber      tenure MonthlyCharges
## 1                 0           1             0 -1.27735389     -1.1602405
## 2                 1           1             0  0.06632271     -0.2596105
## 3                 0           1             0 -1.23663642     -0.3626346
## 4                 1           1             0  0.51421491     -0.7464825
## 5                 0           0             1 -1.23663642      0.1973512
## 6                 0           0             1 -0.99233158      1.1594634
##   TotalCharges
## 1   -0.9941234
## 2   -0.1737275
## 3   -0.9595809
## 4   -0.1952338
## 5   -0.9403906
## 6   -0.6453233
```

```
## 'data.frame':    7043 obs. of  23 variables:
##  $ gender           : Factor w/ 2 levels "0","1": 2 1 1 1 2 2 1 2 2 1 ...
##  $ SeniorCitizen    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Partner          : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Dependents       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 2 ...
##  $ PhoneService     : Factor w/ 2 levels "0","1": 1 2 2 1 2 2 2 1 2 2 ...
##  $ OnlineSecurity   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 2 1 2 ...
##  $ OnlineBackup     : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 2 1 1 2 ...
##  $ DeviceProtection : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 1 2 1 ...
##  $ TechSupport      : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 2 1 ...
##  $ StreamingTV      : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 1 2 1 ...
##  $ StreamingMovies  : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 2 1 ...
##  $ PaperlessBilling : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 1 2 1 ...
##  $ Churn            : Factor w/ 2 levels "0","1": 1 1 2 1 2 2 1 1 2 1 ...
##  $ MultipleLines_no : Factor w/ 2 levels "0","1": 1 2 2 1 2 1 1 1 1 2 ...
##  $ Pay_Mailed       : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 2 1 1 ...
##  $ Pay_Credit       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
##  $ Contract_Month   : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 2 2 1 ...
##  $ Contract_One_year: Factor w/ 2 levels "0","1": 1 2 1 2 1 1 1 1 1 2 ...
##  $ Service_DSL      : Factor w/ 2 levels "0","1": 2 2 2 2 1 1 1 2 1 2 ...
##  $ Service_Fiber    : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 2 1 2 1 ...
##  $ tenure           : num  -1.2774 0.0663 -1.2366 0.5142 -1.2366 ...
##  $ MonthlyCharges   : num  -1.16 -0.26 -0.363 -0.746 0.197 ...
##  $ TotalCharges     : num  -0.994 -0.174 -0.96 -0.195 -0.94 ...
```

로지스틱 회귀분석

```
# 로지스틱 회귀분석에는 glm 함수가 이용 된다. glm 함수 선형적이지 않은 대상을 선형적으로 일반화 시킨 모형으로, 종속변수 및 독립변수를 차례대로 입력하고 사이에 '~' 기호를 삽입해준다.
# 종속변수는 Churn이고, 독립변수가 gender ~ TotalCharges 까지이며, family = binomial로 이진형태로 지정한 후 summary로 분석을 진행
logi_1 <- glm(Churn ~ gender+SeniorCitizen+Partner+Dependents+PhoneService+OnlineSecurity+OnlineBackup+DeviceProtection+TechSupport+StreamingTV+StreamingMovies+PaperlessBilling+MultipleLines_no+Pay_Mailed+Pay_Credit+Contract_Month+Contract_One_year+Service_DSL+Service_Fiber+tenure+MonthlyCharges+TotalCharges, family = binomial, data = final_telco) ;
 summary(logi_1)
```

```
##
## Call:
## glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents +
##     PhoneService + OnlineSecurity + OnlineBackup + DeviceProtection +
##     TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
##     MultipleLines_no + Pay_Mailed + Pay_Credit + Contract_Month +
##     Contract_One_year + Service_DSL + Service_Fiber + tenure +
##     MonthlyCharges + TotalCharges, family = binomial, data = final_telco)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9060  -0.6805  -0.2860   0.7354   3.4605
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.65456    2.06657  -2.736  0.00622 **
## gender1             0.01963    0.06471   0.303  0.76157
## SeniorCitizen1      0.22601    0.08431   2.681  0.00735 **
## Partner1           -0.00374    0.07767  -0.048  0.96160
## Dependents1        -0.15144    0.08967  -1.689  0.09124 .
## PhoneService1       0.60213    0.80472   0.748  0.45431
## OnlineSecurity1    -0.21780    0.17833  -1.221  0.22196
## OnlineBackup1       0.02435    0.17500   0.139  0.88933
## DeviceProtection1   0.14133    0.17606   0.803  0.42213
## TechSupport1       -0.19120    0.18020  -1.061  0.28867
## StreamingTV1        0.60009    0.32570   1.842  0.06540 .
## StreamingMovies1    0.60781    0.32601   1.864  0.06227 .
## PaperlessBilling1   0.35039    0.07435   4.712 2.45e-06 ***
## MultipleLines_no1  -0.44891    0.17703  -2.536  0.01122 *
## Pay_Mailed1        -0.27397    0.09269  -2.956  0.00312 **
## Pay_Credit1        -0.29761    0.09319  -3.193  0.00141 **
## Contract_Month1     1.38796    0.17622   7.876 3.37e-15 ***
## Contract_One_year1  0.71027    0.17656   4.023 5.75e-05 ***
## Service_DSL1        1.80317    0.80580   2.238  0.02524 *
## Service_Fiber1      3.56613    1.59069   2.242  0.02497 *
## tenure             -1.51961    0.15315  -9.923  < 2e-16 ***
## MonthlyCharges     -1.21446    0.95371  -1.273  0.20287
## TotalCharges        0.75658    0.16033   4.719 2.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8143.4  on 7031  degrees of freedom
## Residual deviance: 5836.7  on 7009  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 5882.7
##
## Number of Fisher Scoring iterations: 6
```

```
# 회귀계수 : coef , Odd ratio : exp
coef(logi_1) ; exp(coef(logi_1))
```

```
##        (Intercept)            gender1     SeniorCitizen1           Partner1
##       -5.654563127         0.019635156         0.226011380        -0.003739645
##         Dependents1      PhoneService1    OnlineSecurity1      OnlineBackup1
##       -0.151437280         0.602131870        -0.217804903         0.024351231
##   DeviceProtection1       TechSupport1       StreamingTV1   StreamingMovies1
##        0.141331739        -0.191204358         0.600088653         0.607814139
##   PaperlessBilling1  MultipleLines_no1         Pay_Mailed1         Pay_Credit1
##        0.350389691        -0.448909008        -0.273968624        -0.297606376
##     Contract_Month1 Contract_One_year1       Service_DSL1     Service_Fiber1
##        1.387959020         0.710267323         1.803166519         3.566133525
##              tenure     MonthlyCharges       TotalCharges
##       -1.519607234        -1.214461110         0.756577858
```

```
##        (Intercept)            gender1     SeniorCitizen1           Partner1
##        0.003501502         1.019829194         1.253589931         0.996267338
##         Dependents1      PhoneService1    OnlineSecurity1      OnlineBackup1
##        0.859471786         1.826007465         0.804282339         1.024650144
##   DeviceProtection1       TechSupport1       StreamingTV1   StreamingMovies1
##        1.151806684         0.825963778         1.822280343         1.836412866
##   PaperlessBilling1  MultipleLines_no1         Pay_Mailed1         Pay_Credit1
##        1.419620655         0.638324178         0.760355932         0.742593585
##     Contract_Month1 Contract_One_year1       Service_DSL1     Service_Fiber1
##        4.006664182         2.034535063         6.068834150        35.379534255
##              tenure     MonthlyCharges       TotalCharges
##        0.218797806         0.296869952         2.130971242
```

- 분석 결과 모든 독립변수가 유의한 값을 가지지 않았기 때문에 회귀모형을 재정의할 필요가 있었음.

```
# 유의하지 않은 변수들을 누락하고 다시 로지스틱 회귀모형을 정의하기로 함(후진선택법 사용)
logi_2 <- step(logi_1, direction = "backward") ; summary(logi_2)
```

```
## Start:  AIC=5882.74
## Churn ~ gender + SeniorCitizen + Partner + Dependents + PhoneService +
##     OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + MultipleLines_no +
##     Pay_Mailed + Pay_Credit + Contract_Month + Contract_One_year +
##     Service_DSL + Service_Fiber + tenure + MonthlyCharges + TotalCharges
##
##                     Df Deviance    AIC
## - Partner            1   5836.7 5880.7
## - OnlineBackup       1   5836.8 5880.8
## - gender             1   5836.8 5880.8
## - PhoneService       1   5837.3 5881.3
## - DeviceProtection   1   5837.4 5881.4
## - TechSupport        1   5837.9 5881.9
## - OnlineSecurity     1   5838.2 5882.2
## - MonthlyCharges     1   5838.4 5882.4
## <none>                   5836.7 5882.7
## - Dependents         1   5839.6 5883.6
## - StreamingTV        1   5840.1 5884.1
## - StreamingMovies    1   5840.2 5884.2
## - Service_DSL        1   5841.7 5885.7
## - Service_Fiber      1   5841.8 5885.8
## - MultipleLines_no   1   5843.2 5887.2
## - SeniorCitizen      1   5843.9 5887.9
## - Pay_Mailed         1   5845.5 5889.5
## - Pay_Credit         1   5847.1 5891.1
## - Contract_One_year  1   5854.4 5898.4
## - PaperlessBilling   1   5859.1 5903.1
## - TotalCharges       1   5860.2 5904.2
## - Contract_Month     1   5911.3 5955.3
## - tenure             1   5953.6 5997.6
##
## Step:  AIC=5880.74
## Churn ~ gender + SeniorCitizen + Dependents + PhoneService +
##     OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + MultipleLines_no +
##     Pay_Mailed + Pay_Credit + Contract_Month + Contract_One_year +
##     Service_DSL + Service_Fiber + tenure + MonthlyCharges + TotalCharges
##
##                     Df Deviance    AIC
## - OnlineBackup       1   5836.8 5878.8
## - gender             1   5836.8 5878.8
## - PhoneService       1   5837.3 5879.3
## - DeviceProtection   1   5837.4 5879.4
## - TechSupport        1   5837.9 5879.9
## - OnlineSecurity     1   5838.2 5880.2
## - MonthlyCharges     1   5838.4 5880.4
## <none>                   5836.7 5880.7
## - StreamingTV        1   5840.1 5882.1
## - StreamingMovies    1   5840.2 5882.2
## - Dependents         1   5840.3 5882.3
## - Service_DSL        1   5841.7 5883.7
## - Service_Fiber      1   5841.8 5883.8
```

```
## - MultipleLines_no   1   5843.2 5885.2
## - SeniorCitizen      1   5844.0 5886.0
## - Pay_Mailed         1   5845.5 5887.5
## - Pay_Credit         1   5847.1 5889.1
## - Contract_One_year  1   5854.4 5896.4
## - PaperlessBilling   1   5859.1 5901.1
## - TotalCharges       1   5860.2 5902.2
## - Contract_Month     1   5911.3 5953.3
## - tenure             1   5954.6 5996.6
##
## Step:  AIC=5878.76
## Churn ~ gender + SeniorCitizen + Dependents + PhoneService +
##     OnlineSecurity + DeviceProtection + TechSupport + StreamingTV +
##     StreamingMovies + PaperlessBilling + MultipleLines_no + Pay_Mailed +
##     Pay_Credit + Contract_Month + Contract_One_year + Service_DSL +
##     Service_Fiber + tenure + MonthlyCharges + TotalCharges
##
##                     Df Deviance    AIC
## - gender             1   5836.9 5876.9
## - DeviceProtection   1   5838.1 5878.1
## - PhoneService       1   5838.6 5878.6
## <none>                   5836.8 5878.8
## - Dependents         1   5840.3 5880.3
## - TechSupport        1   5840.4 5880.4
## - OnlineSecurity     1   5841.5 5881.5
## - MonthlyCharges     1   5843.5 5883.5
## - SeniorCitizen      1   5844.0 5884.0
## - Pay_Mailed         1   5845.5 5885.5
## - Pay_Credit         1   5847.1 5887.1
## - StreamingTV        1   5849.1 5889.1
## - StreamingMovies    1   5849.6 5889.6
## - MultipleLines_no   1   5853.3 5893.3
## - Contract_One_year  1   5854.4 5894.4
## - Service_DSL        1   5855.8 5895.8
## - Service_Fiber      1   5858.4 5898.4
## - PaperlessBilling   1   5859.2 5899.2
## - TotalCharges       1   5860.2 5900.2
## - Contract_Month     1   5911.3 5951.3
## - tenure             1   5954.6 5994.6
##
## Step:  AIC=5876.86
## Churn ~ SeniorCitizen + Dependents + PhoneService + OnlineSecurity +
##     DeviceProtection + TechSupport + StreamingTV + StreamingMovies +
##     PaperlessBilling + MultipleLines_no + Pay_Mailed + Pay_Credit +
##     Contract_Month + Contract_One_year + Service_DSL + Service_Fiber +
##     tenure + MonthlyCharges + TotalCharges
##
##                     Df Deviance    AIC
## - DeviceProtection   1   5838.2 5876.2
## - PhoneService       1   5838.7 5876.7
## <none>                   5836.9 5876.9
## - Dependents         1   5840.4 5878.4
## - TechSupport        1   5840.5 5878.5
## - OnlineSecurity     1   5841.6 5879.6
```

```
## - MonthlyCharges     1   5843.6 5881.6
## - SeniorCitizen      1   5844.1 5882.1
## - Pay_Mailed         1   5845.7 5883.7
## - Pay_Credit         1   5847.2 5885.2
## - StreamingTV        1   5849.2 5887.2
## - StreamingMovies    1   5849.6 5887.6
## - MultipleLines_no   1   5853.4 5891.4
## - Contract_One_year  1   5854.5 5892.5
## - Service_DSL        1   5855.8 5893.8
## - Service_Fiber      1   5858.5 5896.5
## - PaperlessBilling   1   5859.3 5897.3
## - TotalCharges       1   5860.3 5898.3
## - Contract_Month     1   5911.4 5949.4
## - tenure             1   5954.8 5992.8
##
## Step:  AIC=5876.18
## Churn ~ SeniorCitizen + Dependents + PhoneService + OnlineSecurity +
##     TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
##     MultipleLines_no + Pay_Mailed + Pay_Credit + Contract_Month +
##     Contract_One_year + Service_DSL + Service_Fiber + tenure +
##     MonthlyCharges + TotalCharges
##
##                     Df Deviance    AIC
## - PhoneService       1   5838.8 5874.8
## <none>                   5838.2 5876.2
## - Dependents         1   5841.8 5877.8
## - MonthlyCharges     1   5844.1 5880.1
## - TechSupport        1   5844.8 5880.8
## - SeniorCitizen      1   5845.4 5881.4
## - OnlineSecurity     1   5846.9 5882.9
## - Pay_Mailed         1   5846.9 5882.9
## - Pay_Credit         1   5848.4 5884.4
## - StreamingTV        1   5849.9 5885.9
## - StreamingMovies    1   5850.5 5886.5
## - MultipleLines_no   1   5853.9 5889.9
## - Contract_One_year  1   5855.6 5891.6
## - Service_DSL        1   5859.1 5895.1
## - PaperlessBilling   1   5860.4 5896.4
## - TotalCharges       1   5861.5 5897.5
## - Service_Fiber      1   5864.0 5900.0
## - Contract_Month     1   5911.8 5947.8
## - tenure             1   5956.3 5992.3
##
## Step:  AIC=5874.82
## Churn ~ SeniorCitizen + Dependents + OnlineSecurity + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + MultipleLines_no +
##     Pay_Mailed + Pay_Credit + Contract_Month + Contract_One_year +
##     Service_DSL + Service_Fiber + tenure + MonthlyCharges + TotalCharges
##
##                     Df Deviance    AIC
## <none>                   5838.8 5874.8
## - Dependents         1   5842.5 5876.5
## - SeniorCitizen      1   5846.0 5880.0
## - Pay_Mailed         1   5847.6 5881.6
```

```
## - Pay_Credit         1   5849.2 5883.2
## - TechSupport        1   5850.1 5884.1
## - MonthlyCharges     1   5851.4 5885.4
## - OnlineSecurity     1   5852.7 5886.7
## - StreamingTV        1   5855.5 5889.5
## - Contract_One_year  1   5856.3 5890.3
## - StreamingMovies    1   5856.6 5890.6
## - MultipleLines_no   1   5859.1 5893.1
## - PaperlessBilling   1   5861.1 5895.1
## - TotalCharges       1   5861.8 5895.8
## - Service_DSL        1   5891.5 5925.5
## - Service_Fiber      1   5906.2 5940.2
## - Contract_Month     1   5913.3 5947.3
## - tenure             1   5957.3 5991.3
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Dependents + OnlineSecurity +
##     TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
##     MultipleLines_no + Pay_Mailed + Pay_Credit + Contract_Month +
##     Contract_One_year + Service_DSL + Service_Fiber + tenure +
##     MonthlyCharges + TotalCharges, family = binomial, data = final_telco)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9037  -0.6797  -0.2859   0.7382   3.4624
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.18444    0.29807 -14.038  < 2e-16 ***
## SeniorCitizen1      0.22436    0.08374   2.679 0.007377 **
## Dependents1        -0.15446    0.08137  -1.898 0.057670 .
## OnlineSecurity1    -0.32718    0.08825  -3.708 0.000209 ***
## TechSupport1       -0.30049    0.08958  -3.354 0.000795 ***
## StreamingTV1        0.38404    0.09429   4.073 4.64e-05 ***
## StreamingMovies1    0.39217    0.09319   4.208 2.57e-05 ***
## PaperlessBilling1   0.34946    0.07430   4.704 2.56e-06 ***
## MultipleLines_no1  -0.32624    0.07275  -4.485 7.30e-06 ***
## Pay_Mailed1        -0.27334    0.09260  -2.952 0.003158 **
## Pay_Credit1        -0.29755    0.09312  -3.195 0.001397 **
## Contract_Month1     1.38282    0.17582   7.865 3.69e-15 ***
## Contract_One_year1  0.70654    0.17650   4.003 6.26e-05 ***
## Service_DSL1        1.22516    0.17131   7.152 8.58e-13 ***
## Service_Fiber1      2.45857    0.30060   8.179 2.87e-16 ***
## tenure             -1.52210    0.15241  -9.987  < 2e-16 ***
## MonthlyCharges     -0.55183    0.15591  -3.539 0.000401 ***
## TotalCharges        0.74706    0.15981   4.675 2.94e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8143.4  on 7031  degrees of freedom
## Residual deviance: 5838.8  on 7014  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 5874.8
##
## Number of Fisher Scoring iterations: 6
```

```
coef(logi_2) ; exp(coef(logi_2))
```

```
##        (Intercept)     SeniorCitizen1        Dependents1    OnlineSecurity1
##         -4.1844431          0.2243568         -0.1544595         -0.3271814
##       TechSupport1       StreamingTV1   StreamingMovies1  PaperlessBilling1
##         -0.3004938          0.3840359          0.3921721          0.3494622
##  MultipleLines_no1        Pay_Mailed1        Pay_Credit1    Contract_Month1
##         -0.3262444         -0.2733373         -0.2975452          1.3828178
## Contract_One_year1       Service_DSL1     Service_Fiber1             tenure
##          0.7065407          1.2251575          2.4585658         -1.5220980
##     MonthlyCharges       TotalCharges
##         -0.5518319          0.7470583
```

```
##        (Intercept)     SeniorCitizen1        Dependents1    OnlineSecurity1
##         0.01523069         1.25151752         0.85687819         0.72095294
##       TechSupport1       StreamingTV1   StreamingMovies1  PaperlessBilling1
##         0.74045248         1.46819820         1.48019240         1.41830452
##  MultipleLines_no1        Pay_Mailed1        Pay_Credit1    Contract_Month1
##         0.72162883         0.76083612         0.74263902         3.98611771
## Contract_One_year1       Service_DSL1     Service_Fiber1             tenure
##         2.02696731         3.40470239        11.68803626         0.21825351
##     MonthlyCharges       TotalCharges
##         0.57589388         2.11078156
```

```
anova(logi_1, logi_2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ gender + SeniorCitizen + Partner + Dependents + PhoneService +
##     OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + MultipleLines_no +
##     Pay_Mailed + Pay_Credit + Contract_Month + Contract_One_year +
##     Service_DSL + Service_Fiber + tenure + MonthlyCharges + TotalCharges
## Model 2: Churn ~ SeniorCitizen + Dependents + OnlineSecurity + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + MultipleLines_no +
##     Pay_Mailed + Pay_Credit + Contract_Month + Contract_One_year +
##     Service_DSL + Service_Fiber + tenure + MonthlyCharges + TotalCharges
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      7009     5836.7
## 2      7014     5838.8 -5  -2.0854   0.8372
```

- 후진선택법을 사용해 유의한 변수만 걸러내도록 하였으며 로지스틱 회귀분석의 결과 해석은 독립변수가 1unit 증가할 때 관심 사건의 log-odd가 증가하는 것이므로 계수를 지수화(Odd ratio)한다. 따라서 각 독립변수들이 1 증가할 때마다 odd ratio가 e^coeff 배 증가한다고 해석해야 한다.
- 따라서 예로 OnlineSecurity(온라인 보안 서비스)의 고객이 1 늘어날 때마다 이탈 확률이 0.72095294배로 줄어들며, Pay_mailed(메일로 결제)를 하는 고객이 1 늘어날 때마다 이탈 확률이 0.76083612배로 줄어든다고 해석한다.
- LRT(Likelihood-ratio test)는 두 모델 간의 적합도를 비교하기 위해 Proposed model이 Null model보다 몇 배나 발생 가능성이 높은지(LR)을 찾아 이를 통해 p값을 계산해 유의성을 판단하는 것으로, 기존 모델과 후진선택으로 만든 모델의 비교 결과 P 값은 0.05보다 큰 0.8372가 나왔기 때문에 유의성이 없다고 판단된다.

```
logi_3 <- glm(Churn ~ SeniorCitizen+OnlineSecurity+TechSupport+StreamingTV+StreamingMovies+PaperlessBilling+MultipleLines_no+Pay_Mailed+Pay_Credit+Contract_Month+Contract_One_year+Service_DSL+tenure+MonthlyCharges+TotalCharges, family = binomial, data = final_telco) ; summary(logi_1)
```

```
##
## Call:
## glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents +
##     PhoneService + OnlineSecurity + OnlineBackup + DeviceProtection +
##     TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
##     MultipleLines_no + Pay_Mailed + Pay_Credit + Contract_Month +
##     Contract_One_year + Service_DSL + Service_Fiber + tenure +
##     MonthlyCharges + TotalCharges, family = binomial, data = final_telco)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9060  -0.6805  -0.2860   0.7354   3.4605
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.65456    2.06657  -2.736  0.00622 **
## gender1             0.01963    0.06471   0.303  0.76157
## SeniorCitizen1      0.22601    0.08431   2.681  0.00735 **
## Partner1           -0.00374    0.07767  -0.048  0.96160
## Dependents1        -0.15144    0.08967  -1.689  0.09124 .
## PhoneService1       0.60213    0.80472   0.748  0.45431
## OnlineSecurity1    -0.21780    0.17833  -1.221  0.22196
## OnlineBackup1       0.02435    0.17500   0.139  0.88933
## DeviceProtection1   0.14133    0.17606   0.803  0.42213
## TechSupport1       -0.19120    0.18020  -1.061  0.28867
## StreamingTV1        0.60009    0.32570   1.842  0.06540 .
## StreamingMovies1    0.60781    0.32601   1.864  0.06227 .
## PaperlessBilling1   0.35039    0.07435   4.712 2.45e-06 ***
## MultipleLines_no1  -0.44891    0.17703  -2.536  0.01122 *
## Pay_Mailed1        -0.27397    0.09269  -2.956  0.00312 **
## Pay_Credit1        -0.29761    0.09319  -3.193  0.00141 **
## Contract_Month1     1.38796    0.17622   7.876 3.37e-15 ***
## Contract_One_year1  0.71027    0.17656   4.023 5.75e-05 ***
## Service_DSL1        1.80317    0.80580   2.238  0.02524 *
## Service_Fiber1      3.56613    1.59069   2.242  0.02497 *
## tenure             -1.51961    0.15315  -9.923  < 2e-16 ***
## MonthlyCharges     -1.21446    0.95371  -1.273  0.20287
## TotalCharges        0.75658    0.16033   4.719 2.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8143.4  on 7031  degrees of freedom
## Residual deviance: 5836.7  on 7009  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 5882.7
##
## Number of Fisher Scoring iterations: 6
```

```
coef(logi_3) ; exp(coef(logi_3))
```

```
##        (Intercept)     SeniorCitizen1    OnlineSecurity1       TechSupport1
##         -2.41461611         0.29445440        -0.49864302        -0.48447280
##       StreamingTV1   StreamingMovies1  PaperlessBilling1  MultipleLines_no1
##          0.04683802         0.07444525         0.39399135        -0.37583985
##        Pay_Mailed1        Pay_Credit1    Contract_Month1 Contract_One_year1
##         -0.35908392        -0.33224774         1.51756498         0.74779823
##       Service_DSL1             tenure     MonthlyCharges       TotalCharges
##          0.04597847        -1.44706630         0.58699570         0.56450671
```

```
##        (Intercept)     SeniorCitizen1    OnlineSecurity1       TechSupport1
##          0.08940165         1.34239375         0.60735427         0.61602188
##       StreamingTV1   StreamingMovies1  PaperlessBilling1  MultipleLines_no1
##          1.04795224         1.07728636         1.48288773         0.68671230
##        Pay_Mailed1        Pay_Credit1    Contract_Month1 Contract_One_year1
##          0.69831575         0.71730959         4.56110529         2.11234400
##       Service_DSL1             tenure     MonthlyCharges       TotalCharges
##          1.04705186         0.23525946         1.79857682         1.75858007
```

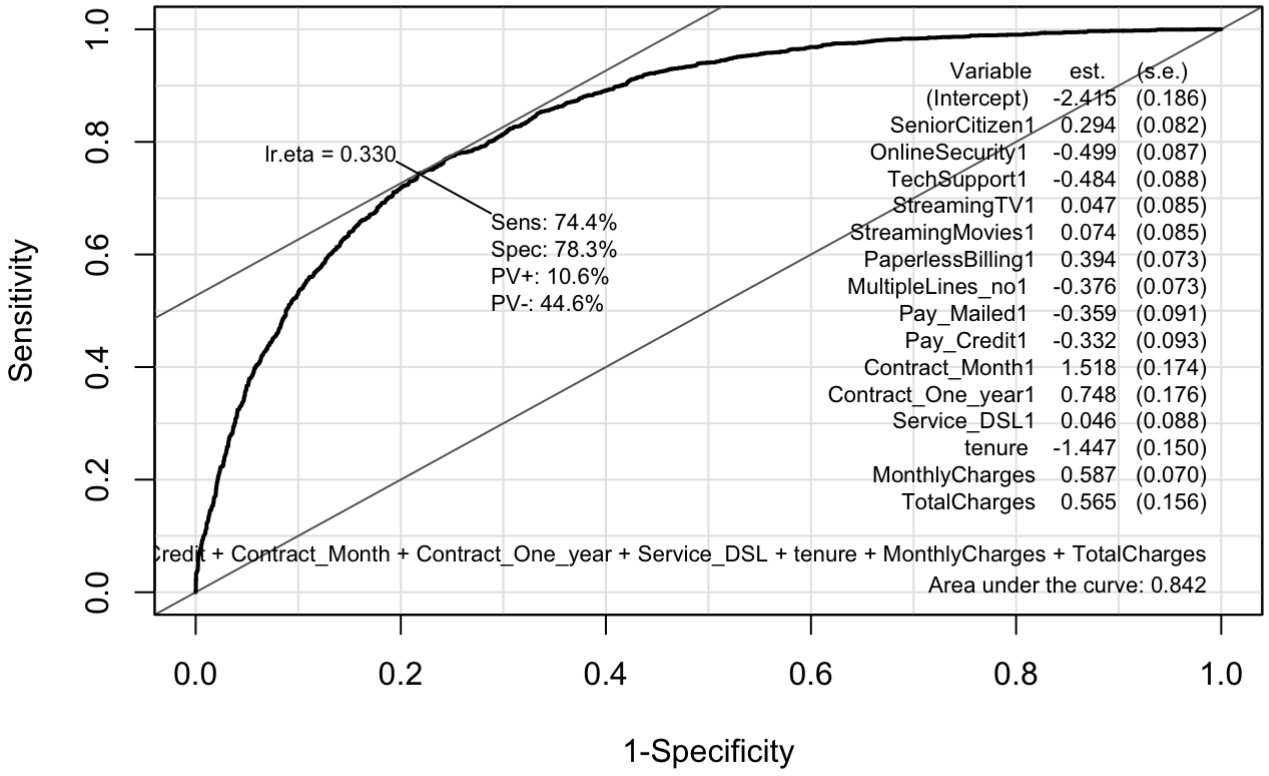
```
anova(logi_1, logi_3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: Churn ~ gender + SeniorCitizen + Partner + Dependents + PhoneService +
##     OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + MultipleLines_no +
##     Pay_Mailed + Pay_Credit + Contract_Month + Contract_One_year +
##     Service_DSL + Service_Fiber + tenure + MonthlyCharges + TotalCharges
## Model 2: Churn ~ SeniorCitizen + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + PaperlessBilling + MultipleLines_no + Pay_Mailed +
##     Pay_Credit + Contract_Month + Contract_One_year + Service_DSL +
##     tenure + MonthlyCharges + TotalCharges
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      7009     5836.7
## 2      7016     5911.2 -7  -74.492 1.819e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 그 이유를 다시 summary(logi_2) 에서 찾아본 결과 유의도가 0.05를 넘어가는 변수가 후진선택법에서 지워지지 않았고 회귀 계수가 비정상적으로 나오는 변수가 출력되는 것을 확인할 수 있었다. 따라서 이들을 소거한 후 다시 logi_3에 대입하여 LR-Test를 진행한 결과 유의미한 값을 가지는 결과를 도출하였다.

ROC 커브를 통해 로지스틱 회귀모델의 성능을 평가

```
# ROC(Receiver Operating Characteristic) : 이진 분류기의 유용성을 검증하는 방식 중 하나
graph_1 <- ROC(form = Churn ~ SeniorCitizen + OnlineSecurity + TechSupport +
    StreamingTV + StreamingMovies + PaperlessBilling + MultipleLines_no +
    Pay_Mailed + Pay_Credit + Contract_Month + Contract_One_year +
    Service_DSL + tenure + MonthlyCharges + TotalCharges, data =final_telco, plot = "RO
C")
```




```
head(graph_1$res); tail(graph_1$res)
```

```
##                      sens         spec pvp       pvn      lr.eta
##                         1 0.0000000000 NaN 0.7342150        -Inf
## 0.001374599547252       1 0.0001936858   0 0.7341772 0.001374600
## 0.00139744629781927     1 0.0003873717   0 0.7341394 0.001397446
## 0.00141143429752969     1 0.0005810575   0 0.7341016 0.001411434
## 0.00142628896469358     1 0.0007747434   0 0.7340637 0.001426289
## 0.00144982012585829     1 0.0009684292   0 0.7340259 0.001449820
```

```
##                           sens spec       pvp pvn    lr.eta
## 0.855861453148219 0.0026752274    1 0.2652626   0 0.8558615
## 0.856468124822018 0.0021401819    1 0.2653671   0 0.8564681
## 0.859119095101767 0.0016051364    1 0.2654716   0 0.8591191
## 0.872206172884255 0.0010700910    1 0.2655761   0 0.8722062
## 0.873572823684446 0.0005350455    1 0.2656806   0 0.8735728
## 0.876586343695934 0.0000000000    1 0.2657850 NaN 0.8765863
```

```
graph_1$res[round(graph_1$res$lr.eta,3) == 0.001,]
```

```
##                      sens         spec pvp       pvn      lr.eta
## 0.001374599547252       1 0.0001936858   0 0.7341772 0.001374600
## 0.00139744629781927     1 0.0003873717   0 0.7341394 0.001397446
## 0.00141143429752969     1 0.0005810575   0 0.7341016 0.001411434
## 0.00142628896469358     1 0.0007747434   0 0.7340637 0.001426289
## 0.00144982012585829     1 0.0009684292   0 0.7340259 0.001449820
## 0.00146621419585402     1 0.0011621150   0 0.7339880 0.001466214
## 0.00146788619595421     1 0.0013558009   0 0.7339502 0.001467886
## 0.00146795694177399     1 0.0015494867   0 0.7339123 0.001467957
## 0.00147074136170153     1 0.0017431726   0 0.7338744 0.001470741
## 0.00147687235338412     1 0.0019368584   0 0.7338365 0.001476872
## 0.00147792726764855     1 0.0021305443   0 0.7337986 0.001477927
## 0.00149458277655067     1 0.0023242301   0 0.7337607 0.001494583
```

```
graph_1$AUC
```

```
## [1] 0.8423857
```

```
graph_1$lr
```

```
##
## Call:  glm(formula = form, family = binomial, data = data)
##
## Coefficients:
##        (Intercept)      SeniorCitizen1     OnlineSecurity1        TechSupport1
##            -2.41462             0.29445            -0.49864            -0.48447
##       StreamingTV1    StreamingMovies1   PaperlessBilling1   MultipleLines_no1
##            0.04684             0.07445             0.39399            -0.37584
##        Pay_Mailed1         Pay_Credit1     Contract_Month1  Contract_One_year1
##           -0.35908            -0.33225             1.51756             0.74780
##       Service_DSL1              tenure      MonthlyCharges        TotalCharges
##            0.04598            -1.44707             0.58700             0.56451
##
## Degrees of Freedom: 7031 Total (i.e. Null);  7016 Residual
##   (11 observations deleted due to missingness)
## Null Deviance:       8143
## Residual Deviance: 5911  AIC: 5943
```

- ROC 커브는 민감도(Sensitivity)와 특이도(Specificity) 간의 관계를 2차원 상의 표현한 것으로 로지스틱 회귀모델의 성능을 평가한다. 해당 그래프를 통해 민감도는 74.4%, 특이도는 78.3%이다. 또한 ROC를 통해 성능을 평가할 때 AUC(Area under the curve)가 중요한 지표가 되는데 1이면 완벽한 모델, 0.5면 random guessing한 모델이라고 한다. AUC가 평균적으로 0.9 이상이면 우수한 모형이라고 하나 해당 그래프에서는 0.842로 그 아래인 점을 확인할 수 있다.
- 또한 최적의 절사점을 보면 lr.eta 값이 0.330으로 나와있는 것을 학인할 수 있는데, 이 값은 일반화 회귀모형에서 도출된 것이다.

```
# 예측 값과 실제 관찰값을 비교하는 과정
# 데이터 수가 몇천개 단위이므로 출력 과정은 생략
# logi_3$data
# logi_3$y
# logi_3$linear.predictors
# cbind(ifelse(1/(1+exp(-logi_3$linear.predictors))>0.001,1,0), logi_3$y)

table(ifelse(1/(1+exp(logi_3$linear.predictors))>0.001,1,0),
      logi_3$y)
```

```
##
##        0    1
##   1 5163 1869
```

- 로지스틱 회귀분석을 통해 예측한 값들과 실제 관찰값을 비교하여 다른 정확도 지표도 계산이 가능하다.
- logi_2 linear.predictors 에서 값(z) 안에서 p 값을 계산했으며, Cut-off value인 0.001을 기준으로 0과 1을 구분하여 실제 관찰값인 logi_2 y 와 비교할 수 있다.

나이브 베이즈 분석

```
# 가설 : 독립변수(고객의 이용 중인 서비스, 계정 관련 정보, 인구 통계학적 정보)에 따라 고객 이탈 여부(종속변수)가
 다를 것이다.

# 데이터는 오직 이진 형태(1,0)로만 사용할 것이기 때문에 수치형 독립변수는 제외
naive_telco <- final_telco[,-c(21,22,23)];head(naive_telco)
```

```
##   gender SeniorCitizen Partner Dependents PhoneService OnlineSecurity
## 1      1             0       1          0            0              0
## 2      0             0       0          0            1              1
## 3      0             0       0          0            1              1
## 4      0             0       0          0            0              1
## 5      1             0       0          0            1              0
## 6      1             0       0          0            1              0
##   OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies
## 1            1                0           0           0               0
## 2            0                1           0           0               0
## 3            1                0           0           0               0
## 4            0                1           1           0               0
## 5            0                0           0           0               0
## 6            0                1           0           1               1
##   PaperlessBilling Churn MultipleLines_no Pay_Mailed Pay_Credit Contract_Month
## 1                1     0                0          0          0              1
## 2                0     0                1          1          0              0
## 3                1     1                1          1          0              1
## 4                0     0                0          0          0              0
## 5                1     1                1          0          0              1
## 6                1     1                0          0          0              1
##   Contract_One_year Service_DSL Service_Fiber
## 1                 0           1             0
## 2                 1           1             0
## 3                 0           1             0
## 4                 1           1             0
## 5                 0           0             1
## 6                 0           0             1
```

```
# 이미 label로 전환할 모든 변수가 factor 형이므로 factor로 변환하는 단계 생략

# 학습 데이터와 평가 데이터를 분할해야함(library는 e1071을 사용)
# train 데이터로 나이브베이즈 모델을 훈련시킨다.
# 학습 데이터는 7:3으로 나눔
bal <- sample(2, nrow(naive_telco), replace = T, prob = c(0.7,0.3))
train_telco <- naive_telco[bal == 1, ]
test_telco <- naive_telco[bal == 2, ]
# 테이블 개수가 가르칠 때마다 바뀌는 것을 알 수 있다. 그러나 데이터의 table 개수보다는 비율을 비교해야 하므로 상관 없음
nrow(train_telco); nrow(test_telco)
```

```
## [1] 4968
```

```
## [1] 2075
```

```
# 전체, 학습 데이터의 Churn 비율 확인
# 두 데이터 모델 전부 비율이 동일함을 확인할 수 있음
print(table(naive_telco$Churn)); print(table(train_telco$Churn))
```

```
##
##    0    1
## 5174 1869
```

```
##
##    0    1
## 3658 1310
```

- 결과 값에 따라 학습 데이터가 출력되는 것을 확인할 수 있으며 또한, 전체 데이터와 학습 데이터의 비율을 비교한 결과 유사하다는 것을 알 수 있었다.

```
naive_1 <- naiveBayes(Churn~., data = train_telco)
naive_1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##         0         1
## 0.7363124 0.2636876
##
## Conditional probabilities:
##    gender
## Y           0         1
##   0 0.4978130 0.5021870
##   1 0.4946565 0.5053435
##
##    SeniorCitizen
## Y           0         1
##   0 0.8696009 0.1303991
##   1 0.7526718 0.2473282
##
##    Partner
## Y           0         1
##   0 0.4696555 0.5303445
##   1 0.6290076 0.3709924
##
##    Dependents
## Y           0         1
##   0 0.6514489 0.3485511
##   1 0.8129771 0.1870229
##
##    PhoneService
## Y            0          1
##   0 0.10005467 0.89994533
##   1 0.09389313 0.90610687
##
##    OnlineSecurity
## Y           0         1
##   0 0.6686714 0.3313286
##   1 0.8335878 0.1664122
##
##    OnlineBackup
## Y           0         1
##   0 0.6262985 0.3737015
##   1 0.7206107 0.2793893
##
##    DeviceProtection
## Y           0         1
##   0 0.6386003 0.3613997
##   1 0.7106870 0.2893130
##
##    TechSupport
```

```
## Y           0         1
##   0 0.6686714 0.3313286
##   1 0.8366412 0.1633588
##
##    StreamingTV
## Y           0         1
##   0 0.6366867 0.3633133
##   1 0.5656489 0.4343511
##
##    StreamingMovies
## Y           0         1
##   0 0.6314926 0.3685074
##   1 0.5641221 0.4358779
##
##    PaperlessBilling
## Y           0         1
##   0 0.4636413 0.5363587
##   1 0.2519084 0.7480916
##
##    MultipleLines_no
## Y           0         1
##   0 0.5051941 0.4948059
##   1 0.5549618 0.4450382
##
##    Pay_Mailed
## Y           0         1
##   0 0.7452160 0.2547840
##   1 0.8366412 0.1633588
##
##    Pay_Credit
## Y           0         1
##   0 0.7493166 0.2506834
##   1 0.8763359 0.1236641
##
##    Contract_Month
## Y           0         1
##   0 0.5779114 0.4220886
##   1 0.1198473 0.8801527
##
##    Contract_One_year
## Y           0          1
##   0 0.74767633 0.25232367
##   1 0.90687023 0.09312977
##
##    Service_DSL
## Y           0         1
##   0 0.6139967 0.3860033
##   1 0.7534351 0.2465649
##
##    Service_Fiber
## Y           0         1
##   0 0.6615637 0.3384363
##   1 0.3137405 0.6862595
```

- 위 결과에서 A-priori probabilities는 사전 확률을 나타내며, Conditional probabilities에서는 예측변수들이 범주형일 경우 P(예측변수|Class)를 나타내지만 수치형일 경우 평균과 표준편차를 나타낸다. 그러나 수치형 변수들은 제외하고 분석을 진행했기 때문에 전자의 경우만 나오는 것을 알 수 있다.

```
# 예측 모델을 통해 정확도 측정
test_telco$naive_1 <- predict(naive_1, newdata = test_telco)

# 모형 평가
temp <- table(test_telco$naive_1, test_telco$Churn, dnn = c("predicted", "actual"))
temp
```

```
##          actual
## predicted    0    1
##         0 1200  172
##         1  316  387
```

- 훈련된 모델을 통해 시험 데이터 분류의 예측치를 구하려면 predict() 함수를 사용하면 된다. 여기서 단순 예측값을 구하려면 type 인자에 아무 것도 설정하지 않지만, 사후확률을 구하려면 type = “raw”로 설정해주면 된다. 또한 table 함수를 이용해 모형에 대한 평가를 진행해보았다.

```
# 모델의 성능 평가
result_summary <- data.frame( modle = "NaiveBayes",
                      accuracy = (temp[1,1] + temp[2,2]) / sum(temp),
                      precision = temp[2,2] / (temp[2,1] + temp[2,2]),
                      recall = temp[2,2] / (temp[1,2] + temp[2,2]))
result_summary <- result_summary %>%
  mutate(F1 = 2 * precision * recall /(precision + recall))

result_summary
```

```
##        modle  accuracy precision    recall        F1
## 1 NaiveBayes 0.7648193 0.5504979 0.6923077 0.6133122
```

- 훈련 단계가 끝난 후 테스트 단계에서 모델의 성능 평가가 이루어진다. accuracy(정확도), precision(정밀도), recall(재현율) 등의 지표를 통해 classifier의 모델 성능을 측정할 수 있다. 여기서 정밀도는 분류 모델이 1로 예측했는데, 실제로 1인지를 살펴본다. 재현율은 실제로 참인 값에 모델이 참이라고 예측한 것의 비율을 나타낸다. 정확도의 경우 두 지표 상에서 참인 값을 참이라고 옳게 예측한 경우에 대한 비율을 나타낸다. 해당 결과에서는 0.76임을 알 수 있다. F1 값은 Precision과 Recall의 조화평균이다. 수식은 mutate 이하이다.

```
# confusionMatrix를 활용한 정확도 측정
train_class <- predict(naive_1, newdata = train_telco)
confusionMatrix(train_class, train_telco$Churn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 3004  494
##          1  654  816
##
##                Accuracy : 0.7689
##                  95% CI : (0.7569, 0.7806)
##     No Information Rate : 0.7363
##     P-Value [Acc > NIR] : 6.788e-08
##
##                   Kappa : 0.4274
##
##  Mcnemar's Test P-Value : 2.696e-06
##
##             Sensitivity : 0.8212
##             Specificity : 0.6229
##          Pos Pred Value : 0.8588
##          Neg Pred Value : 0.5551
##              Prevalence : 0.7363
##          Detection Rate : 0.6047
##    Detection Prevalence : 0.7041
##       Balanced Accuracy : 0.7221
##
##        'Positive' Class : 0
##
```

- confusionMatrix는 혼동행렬 혹은 정오표로 불리고 분류 모델의 학습 성능 평가를 위한 행렬이다. 앞선 단계에서 활용했던 모델 평가와 같이 예측값과 데이터의 실제 값의 발생 빈도를 나열한 것이다. Accuracy가 마찬가지로 나타나며, 추가로 95%의 신뢰구간이 나타난다. 이외에 Sensivity, Specificity 등 다양한 지표가 존재한다다.

```
CrossTable(train_class, train_telco$Churn,
           prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  4968
##
##
##              | actual
##    predicted |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |      3004 |       494 |      3498 |
##              |     0.605 |     0.099 |           |
## -------------|-----------|-----------|-----------|
##            1 |       654 |       816 |      1470 |
##              |     0.132 |     0.164 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      3658 |      1310 |      4968 |
## -------------|-----------|-----------|-----------|
##
##
```

- gmodels 라이브러리 내 CrossTable 함수를 통해 표 형태로 정리하였다.
- 위 함수로도 모델 성능을 평가할 수 있다.

라플라스 추정량을 사용하여 모델 성능 개선

```
naive_2 <- naiveBayes(Churn~., data = train_telco, laplace = 1)
naive_2
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##         0         1
## 0.7363124 0.2636876
##
## Conditional probabilities:
##    gender
## Y           0         1
##   0 0.4978142 0.5021858
##   1 0.4946646 0.5053354
##
##    SeniorCitizen
## Y           0         1
##   0 0.8693989 0.1306011
##   1 0.7522866 0.2477134
##
##    Partner
## Y           0         1
##   0 0.4696721 0.5303279
##   1 0.6288110 0.3711890
##
##    Dependents
## Y           0         1
##   0 0.6513661 0.3486339
##   1 0.8125000 0.1875000
##
##    PhoneService
## Y           0         1
##   0 0.1002732 0.8997268
##   1 0.0945122 0.9054878
##
##    OnlineSecurity
## Y           0         1
##   0 0.6685792 0.3314208
##   1 0.8330793 0.1669207
##
##    OnlineBackup
## Y           0         1
##   0 0.6262295 0.3737705
##   1 0.7202744 0.2797256
##
##    DeviceProtection
## Y           0         1
##   0 0.6385246 0.3614754
##   1 0.7103659 0.2896341
##
##    TechSupport
```

```
## Y           0         1
##   0 0.6685792 0.3314208
##   1 0.8361280 0.1638720
##
##    StreamingTV
## Y           0         1
##   0 0.6366120 0.3633880
##   1 0.5655488 0.4344512
##
##    StreamingMovies
## Y           0         1
##   0 0.6314208 0.3685792
##   1 0.5640244 0.4359756
##
##    PaperlessBilling
## Y           0         1
##   0 0.4636612 0.5363388
##   1 0.2522866 0.7477134
##
##    MultipleLines_no
## Y           0         1
##   0 0.5051913 0.4948087
##   1 0.5548780 0.4451220
##
##    Pay_Mailed
## Y          0        1
##   0 0.745082 0.254918
##   1 0.836128 0.163872
##
##    Pay_Credit
## Y           0         1
##   0 0.7491803 0.2508197
##   1 0.8757622 0.1242378
##
##    Contract_Month
## Y           0         1
##   0 0.5778689 0.4221311
##   1 0.1204268 0.8795732
##
##    Contract_One_year
## Y          0        1
##   0 0.747541 0.252459
##   1 0.906250 0.093750
##
##    Service_DSL
## Y           0         1
##   0 0.6139344 0.3860656
##   1 0.7530488 0.2469512
##
##    Service_Fiber
## Y           0         1
##   0 0.6614754 0.3385246
##   1 0.3140244 0.6859756
```

- 나이브 베이즈 모델을 시행할 때 변수의 양이 많아지면 계산의 어려움이 생긴다. 예를 들어 변수가 15개고 값이 2개만 존재하더라도, 2의 15제곱만큼의 값이 나오기 때문에 비어있는 값을 사용할 때 그냥 사용하면 확률이 0이 되는 경우가 발생한다. 따라서 이를 피하고자 라플라스를 사용한다.

```
test_telco$naive_2 <- predict(naive_2, newdata = test_telco)

temp <- table(test_telco$naive_2, test_telco$Churn, dnn = c("predicted", "actual"))

result_temp <- data.frame(modle = "NaiveBayes+LPLC",
                      accuracy = (temp[1,1] + temp[2,2]) / sum(temp),
                       precision = temp[2,2] / (temp[2,1] + temp[2,2]),
                       recall = temp[2,2] / (temp[1,2] + temp[2,2]))

result_temp <- result_temp %>%
  mutate(F1 = 2 * precision * recall / (precision + recall))

result_sum <- bind_rows(result_summary, result_temp)
result_sum
```

```
##             modle  accuracy precision    recall        F1
## 1      NaiveBayes 0.7648193 0.5504979 0.6923077 0.6133122
## 2 NaiveBayes+LPLC 0.7657831 0.5520685 0.6923077 0.6142857
```

- 나이브베이즈만 사용했을 때와 나이브베이즈에 라플라스 옵션을 추가했을 때를 비교한 것이다. 밑에서 다시 설명하겠으나, 결론적으로 나이브 베이즈만 사용했을 때는 전체 데이터 중에 오분류한 값이 존재했다. 그렇게 많지는 않으나 실제 해당 모델이 서비스에 이용된다면 오분류한 값을 줄이는 것이 중요하기 때문에 더 높은 정확도를 갖게 만들어야 한다. 나이브베이즈만 사용할 때와 라플라스를 함께 사용할 때의 값을 비교해보면 Accuracy(정확도)가 라플라스를 사용했을 때가 더 높다는 것을 알 수 있다.

```
train_class_2 <- predict(naive_2, newdata = train_telco)
confusionMatrix(train_class_2, train_telco$Churn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 3003  495
##          1  655  815
##
##                Accuracy : 0.7685
##                  95% CI : (0.7565, 0.7802)
##     No Information Rate : 0.7363
##     P-Value [Acc > NIR] : 9.72e-08
##
##                   Kappa : 0.4264
##
##  Mcnemar's Test P-Value : 2.75e-06
##
##             Sensitivity : 0.8209
##             Specificity : 0.6221
##          Pos Pred Value : 0.8585
##          Neg Pred Value : 0.5544
##              Prevalence : 0.7363
##          Detection Rate : 0.6045
##    Detection Prevalence : 0.7041
##       Balanced Accuracy : 0.7215
##
##        'Positive' Class : 0
##
```

- 라플라스를 사용하지 않고 만든 confusionMatrix와 비교해보면 알겠지만 표 내의 값이 조금 상이하다는 것을 알 수 있다. 이는 라플라스 상수를 추가했을 때 더 높은 정확도를 갖는 모델로 만들 수 있다는 결론이 된다.

```
CrossTable(train_class_2, train_telco$Churn,
           prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
           dnn = c('predicted', 'actual'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  4968
##
##
##                | actual
##    predicted |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##             0 |      3003 |       495 |      3498 |
##               |     0.604 |     0.100 |           |
## -------------|-----------|-----------|-----------|
##             1 |       655 |       815 |      1470 |
##               |     0.132 |     0.164 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      3658 |      1310 |      4968 |
## -------------|-----------|-----------|-----------|
##
##
```