예측 애널리틱스 과제 3

2014170852 산업경영공학부 조영관

<로지스틱 회귀분석 모델 구축하기>

데이터 파악

Target value인 Y 값은 연체 여부를 나타내는 0과 1의 값이다.

즉 범주형 값이므로 일반 회귀분석이 아닌 로지스틱 회귀분석을 진행해야 한다.

그리고 이 연체 여부에 영향을 미치는 독립변수 X들이 다양하게 존재한다. (총 22개의 독립변수)

전체 변수를 대상으로 로지스틱 회귀분석을 해보았다.

```
Call:
glm(formula = TARGET ~ ., family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8485  -0.4013  -0.2559  -0.1384   4.8019

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.995e-01  1.415e-01   2.822 0.004767 **
AGE           -1.988e-03  1.970e-03  -1.009 0.312978
TOT_LOAN      -1.922e-03  3.946e-04  -4.870 1.12e-06 ***
TOT_LOAN_CRD   9.376e-04  4.725e-04   1.984 0.047230 *
LOAN_BNK      -8.650e-03  6.571e-04 -13.165  < 2e-16 ***
LOAN_CPT       1.779e-02  1.239e-03  14.365  < 2e-16 ***
CRDT_CNT      -4.832e-01  1.334e-02 -36.229  < 2e-16 ***
GUARN_CNT      2.416e-01  3.271e-02   7.387 1.50e-13 ***
INCOME         3.659e-05  8.625e-06   4.242 2.21e-05 ***
LOAN_CRD_CNT  -4.517e-02  3.109e-02  -1.453 0.146327
LATE_RATE      8.588e-04  1.706e-03   0.504 0.614604
LATE_RATE_1Y   2.040e-02  1.220e-03  16.723  < 2e-16 ***
INS_MON_MAX    1.362e-05  2.613e-05   0.521 0.602268
CANCEL_CNT_1Y  4.176e-01  6.708e-02   6.225 4.81e-10 ***
CALL_TIME      1.162e-03  7.580e-04   1.533 0.125168
TEL_COST_MON  -3.478e-04  3.038e-04  -1.145 0.252244
MOBILE_PRICE  -5.311e-04  4.990e-05 -10.642  < 2e-16 ***
SUSP_DAY       3.723e-04  1.135e-04   3.280 0.001039 **
LATE_TEL       4.208e-03  3.116e-04  13.505  < 2e-16 ***
COMB_COMM1    -1.461e-01  4.042e-02  -3.614 0.000301 ***
SEXM           1.280e-02  3.845e-02   0.333 0.739182
PAY_METHODB   -1.972e+00  6.961e-02 -28.327  < 2e-16 ***
PAY_METHODC   -1.648e+00  5.713e-02 -28.848  < 2e-16 ***
PAY_METHODD   -8.375e-01  7.690e-02 -10.890  < 2e-16 ***
JOBB           1.817e-01  7.752e-02   2.344 0.019072 *
JOBC           9.085e-02  7.658e-02   1.186 0.235529
JOBD           6.009e-02  9.597e-02   0.626 0.531277
---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27289  on 43385  degrees of freedom
Residual deviance: 19983  on 43359  degrees of freedom
AIC: 20037

Number of Fisher Scoring iterations: 7
```

결과에서 p-value에 주목해보면 몇몇 독립변수들은 0.05를 초과하는 것을 볼 수 있다.

즉, 베타(계수)가 0이라는 귀무가설을 기각하지 못한다. 따라서 y 값을 충분히 설명하는 변수가 될 수 없다. 이 변수들을 제거하고 다시 iteration을 돌려본다.

총 8개의 독립변수를 제거했다.

```
Call:
glm(formula = TARGET ~ ., family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9363  -0.4015  -0.2561  -0.1387   4.7855

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.432e-01  6.666e-02   6.649 2.96e-11 ***
TOT_LOAN      -1.945e-03  3.933e-04  -4.944 7.64e-07 ***
TOT_LOAN_CRD   9.831e-04  4.709e-04   2.088 0.036839 *
LOAN_BNK      -8.607e-03  6.541e-04 -13.158  < 2e-16 ***
LOAN_CPT       1.774e-02  1.237e-03  14.338  < 2e-16 ***
CRDT_CNT      -4.827e-01  1.329e-02 -36.319  < 2e-16 ***
GUARN_CNT      2.402e-01  3.257e-02   7.374 1.65e-13 ***
INCOME         2.970e-05  7.675e-06   3.869 0.000109 ***
LATE_RATE_1Y   2.068e-02  9.758e-04  21.197  < 2e-16 ***
CANCEL_CNT_1Y  4.189e-01  6.623e-02   6.325 2.54e-10 ***
MOBILE_PRICE  -5.368e-04  4.854e-05 -11.059  < 2e-16 ***
SUSP_DAY       3.911e-04  1.123e-04   3.483 0.000496 ***
LATE_TEL       4.093e-03  2.786e-04  14.691  < 2e-16 ***
COMB_COMM1    -1.472e-01  4.032e-02  -3.650 0.000262 ***
PAY_METHODB   -1.969e+00  6.935e-02 -28.385  < 2e-16 ***
PAY_METHODC   -1.654e+00  5.689e-02 -29.080  < 2e-16 ***
PAY_METHODD   -8.417e-01  7.673e-02 -10.969  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value 값을 확인해보면 0.05를 넘는 변수가 없음을 확인할 수 있다.

따라서, 위 14개의 독립변수 x들이 Y 값을 유의미하게 설명한다.

Estimate을 보면 양의 값을 가지는 계수는 그 변수가 한 단위 증가하면 연체 발생의 확률도 증가한다는 것을 의미한다. 반대로 음의 값을 가지는 것은 변수가 한 단위 커질수록 연체 발생 확률이 감소한다. 즉, 신용대출 총액, 카드사 대출 총액, 보증 건수, 소득, 1년 보험료 연체율, 1년 실효해지건수, 회선 사용정지일 수, 핸드폰 요금 연체 총액이 연체 여부 발생의 확률을 높이는데 기여한다. 반면에 대출 총액, 은행권 대출 총액, 신용카드 발급 수, 핸드폰 단말기 가격이 증가함에 따라 연체 여부 발생 확률은 낮아진다. 범주형 변수의 경우, 결합상품을 가입했다면 연체 여부 발생 확률이 낮아지며, 핸드폰 요금 납부방법이 B, C, D 순으로 연체 여부 발생 확률이 낮아진다.

제일 유의미하게 영향을 미치는 변수는 지불 방법(PAY_METHOD), 최근 1년 실효해지건수 (CANCEL_CNT_1Y), 신용카드 발급수(CRDT_CNT)임을 확인할 수 있다.

<Training set 과 Test set으로 나눈 후 Testing data 정확도 계산>

Train set과 test set을 7 대 3으로 나누어 train set을 이용해 로지스틱 회귀분석 모델을 구축했다.

그리고 test data를 통해 모델을 검증하였다.

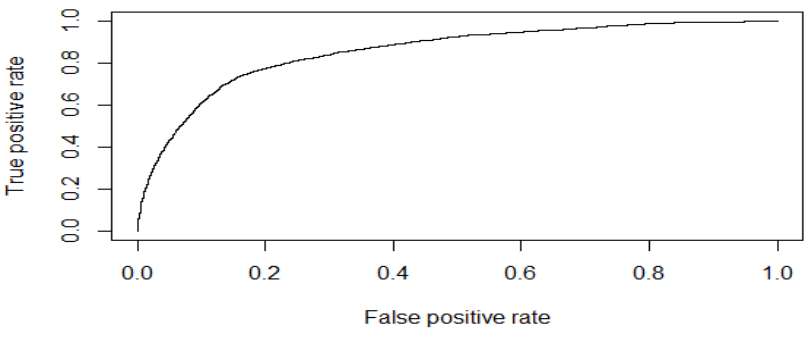
ANOVA 분석을 진행하여 Model의 summary를 확인해보았다.

```
> anova(train_model,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: TARGET

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         30369      18897
TOT_LOAN      1   586.31     30368      18311 < 2.2e-16 ***
TOT_LOAN_CRD  1   126.07     30367      18185 < 2.2e-16 ***
LOAN_BNK      1   500.44     30366      17684 < 2.2e-16 ***
LOAN_CPT      1    48.35     30365      17636 3.574e-12 ***
CRDT_CNT      1  1685.44     30364      15950 < 2.2e-16 ***
GUARN_CNT     1    39.05     30363      15912 4.140e-10 ***
INCOME        1     6.19     30362      15905  0.012865 *
LATE_RATE_1Y  1   414.25     30361      15491 < 2.2e-16 ***
CANCEL_CNT_1Y 1    34.20     30360      15457 4.973e-09 ***
MOBILE_PRICE  1    89.24     30359      15368 < 2.2e-16 ***
SUSP_DAY      1     7.80     30358      15360  0.005233 **
LATE_TEL      1   839.10     30357      14521 < 2.2e-16 ***
COMB_COMM     1    22.37     30356      14498 2.247e-06 ***
PAY_METHOD    3   697.09     30353      13801 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

p-value를 확인해보면 0.05를 넘는 독립변수가 없음을 확인할 수 있다.

위 14개의 독립변수가 Y인 연체 여부에 유의미한 영향을 미친다.

모델 예측 정확도를 확인해보기 위해 ROC CURVE를 통해 시각적으로 확인해보자.



왼쪽 상단으로 곡선이 치우칠수록 모델의 예측 정확도가 높다. 즉 위 CURVE를 살펴보면 비교적 예측 정확도가 높아 보인다. 그 값을 구해보자.

그 수치인 AUC 값을 계산해보면 0.8544가 나온다.

즉, 모델의 예측 정확도는 약 85.44% 이다.

<CUT – OFF 값을 다르게 함에 따라 변화하는 예측 정확도 계산>

```
> binding <- test_data %>% select(TARGET) %>% mutate(per=temp$p)
> binding <- binding %>% mutate(check=ifelse(per>=0.5, 1, 0))
> binding <- binding %>% mutate(last=ifelse(TARGET==check, 0, 1))
> sum(binding$last)
[1] 1181
> binding <- test_data %>% select(TARGET) %>% mutate(per=temp$p)
> binding <- binding %>% mutate(check=ifelse(per>=0.6, 1, 0))
> binding <- binding %>% mutate(last=ifelse(TARGET==check, 0, 1))
> sum(binding$last)
[1] 1177
> binding <- test_data %>% select(TARGET) %>% mutate(per=temp$p)
> binding <- binding %>% mutate(check=ifelse(per>=0.7, 1, 0))
> binding <- binding %>% mutate(last=ifelse(TARGET==check, 0, 1))
> sum(binding$last)
[1] 1220
> binding <- test_data %>% select(TARGET) %>% mutate(per=temp$p)
> binding <- binding %>% mutate(check=ifelse(per>=0.4, 1, 0))
> binding <- binding %>% mutate(last=ifelse(TARGET==check, 0, 1))
> sum(binding$last)
[1] 1195
> binding <- test_data %>% select(TARGET) %>% mutate(per=temp$p)
> binding <- binding %>% mutate(check=ifelse(per>=0.3, 1, 0))
> binding <- binding %>% mutate(last=ifelse(TARGET==check, 0, 1))
> sum(binding$last)
[1] 1248
```

위 코드는 각각 CUT-OFF를 다르게 함에 따라 예측값과 실제 결과값이 다른 값의 개수이다.

즉, CUT-OFF가 0.5일 때 1181개의 값이 예측값과 실제 결과값이 다르게 측정되었다.

이를 활용하여 CUT-OFF가 변화함에 따라 예측 정확도가 어떻게 변화하는지 계산해보았다.

```
> # cut-off 0.5
> 1-1181/13016
[1] 0.9092655
> # cut-off 0.6   -> best
> 1-1177/13016
[1] 0.9095728
> # cut-off 0.7
> 1-1220/13016
[1] 0.9062692
> # cut-off 0.3
> 1-1248/13016
[1] 0.904118
> # cut-off 0.4
> 1-1195/13016
[1] 0.9081899
```

대체로 CUT-OFF를 조정해도 예측 정확도는 거의 유사하지만, 그래도 그 중 가장 좋은 것을 고르면 CUT-OFF가 0.6일 때 예측 정확도가 높고, 제일 적절함을 알 수 있다.

<ODDS 예시>

구글을 통해 검색해보니, Odds 라는 개념이 주로 사용되는 분야가 질병, 의약 분야임을 확인할 수 있었습니다. 약을 복용한 그룹의 odds/약을 복용하지 않은 그룹의 odds를 계산하여 약이 실제로 약효가 있는지를 확인하는 예시가 있습니다.

또는, 어떤 질병이 있을 때의 odds/질병이 없을 때의 odds를 계산하여 환자-대조군 연구에도 사용합니다.

그리고 질병/의약 분야가 아닌 도박/베팅 분야에서도 이 개념이 사용됩니다.

카지노에서 게임을 할 때 유불리를 따질 때 odds를 활용하는 것, 스포츠 베팅을 할 때 팀의 승산이 높고 낮음에 따라 배당률이 달라지는 것에서 odds를 활용함을 알 수 있습니다.

참고자료

https://dermabae.tistory.com/185