고객 이탈 예측을 위한 생존분석

김희영

2023-12-14

탐구 동기 및 목표

- 강의 시간에는 의학 분야 예제로 생존분석을 공부했는데, 기존에 관심있던 마케팅 분야에서 생존분석이 어떻게 활용될 수 있는지 탐구하고자 함.
- 고객 데이터의 생존함수와 위험함수로부터 데이터에 대한 이해를 넓히고자 하며, 이후 Cox 비례 위험 모형을 적합한 후 모형의 예측 정도를 확인하고자 함.

데이터 설명

`liver` 패키지의 `churnTel` 데이터에는 7,043개의 관측치에 대한 21개 변수가 존재함 (Mohammadi and Burke 2023).

변수	설명
customer.ID	Customer ID.
gender	Whether the customer is a male or a female.
senior.citizen	Whether the customer is a senior citizen or not (1, 0).
partner	Whether the customer has a partner or not (yes, no).
dependent	Whether the customer has dependents or not (yes, no).
tenure	Number of months the customer has stayed with the company.
phone.service	Whether the customer has a phone service or not (yes, no).

데이터 설명

변수	설명
multiple.lines	Whether the customer has multiple lines or not (yes, no, no phone service).
internet.service	Customer’s internet service provider (DSL, fiber optic, no).
online.security	Whether the customer has online security or not (yes, no, no internet service).
online.backup	Whether the customer has online backup or not (yes, no, no internet service).
device.protection	Whether the customer has device protection or not (yes, no, no internet service).
tech.support	Whether the customer has tech support or not (yes, no, no internet service).
streaming.TV	Whether the customer has streaming TV or not (yes, no, no internet service).
streaming.movie	Whether the customer has streaming movies or not (yes, no, no internet service).

데이터 설명

변수	설명
contract	The contract term of the customer (month to month, 1 year, 2 year).
paperless.bill	Whether the customer has paperless billing or not (yes, no).
payment.method	The customer's payment method (electronic check, mail check, bank transfer, credit card).
monthly.charge	The amount charged to the customer monthly.
total.charges	The total amount charged to the customer.
churn	Whether the customer churned or not (yes or no).

문제에서 관심있는 event인 `churn`은 고객에 대해서 한 번만 발생하는 event로, 그 값은 `yes` 또는 `no` 임.

비모수적 생존함수 추정 (카플란-마이어 방법)

데이터로부터 생존곡선을 산출하는 Kaplan-Meier 방법을 적용함.

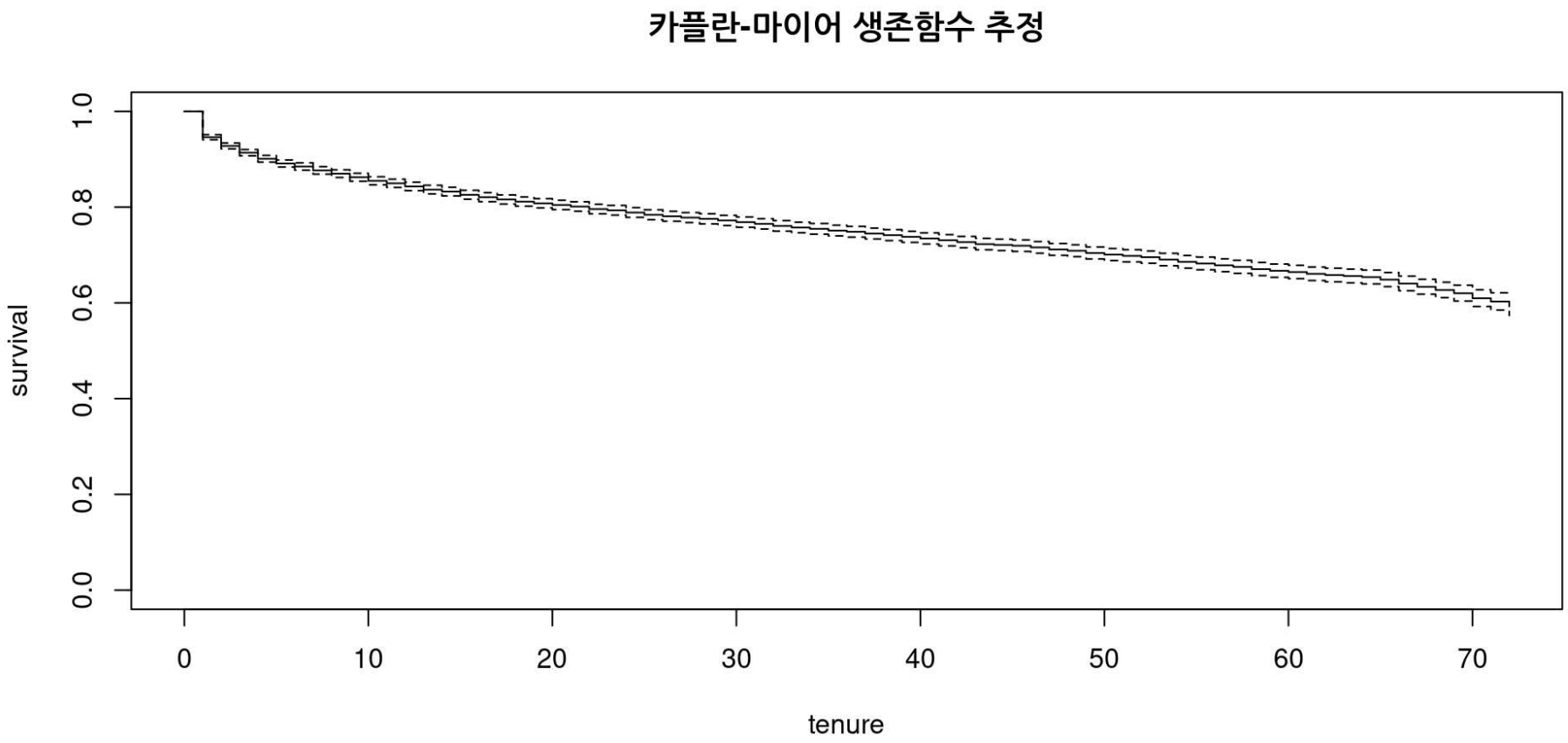
만약 관측치가 censored 된 경우는 0, 그렇지 않으면 1로 status indicator를 정의함. 예제에서 censored 되는 경우는 실험이 종료될 때까지 event가 발생하지 않은 경우로, 다른 censoring의 경우인 loss to follow-up은 해당 데이터에서 발생하지 않음.

```
time <- churnTel$tenure
status <- ifelse(churnTel$churn == "yes", 1, 0)

surv.fit <- survfit(Surv(time, status) ~ 1)

plot(surv.fit, main="카플란-마이어 생존함수 추정",
     xlab="tenure", ylab="survival")
```

비모수적 생존함수 추정 (카플란-마이어 방법)



비모수적 생존함수 추정 (카플란-마이어 방법)

```
output <- tibble(time=surv.fit$time,
                 surv=surv.fit$surv) %>%
  mutate(space=time*surv) %>%
  summarise(avg_life_time=mean(space)) %>%
  pull()


print(output)
```

```
[1] 25.68115
```

생존함수에서 구한 평균이 실제 구독 기간의 평균을 의미하는 것은 아님. 하지만 해당 추정치는 고객 유지 예산 예산 배정의 문제에 도움을 줄 수 있음(Linoff and Berry 2011).

Hazard의 경향 확인

time t에 대한 Hazard는

$$\frac{\#\text{ of event at time t}}{\text{population at risk at time t}}$$

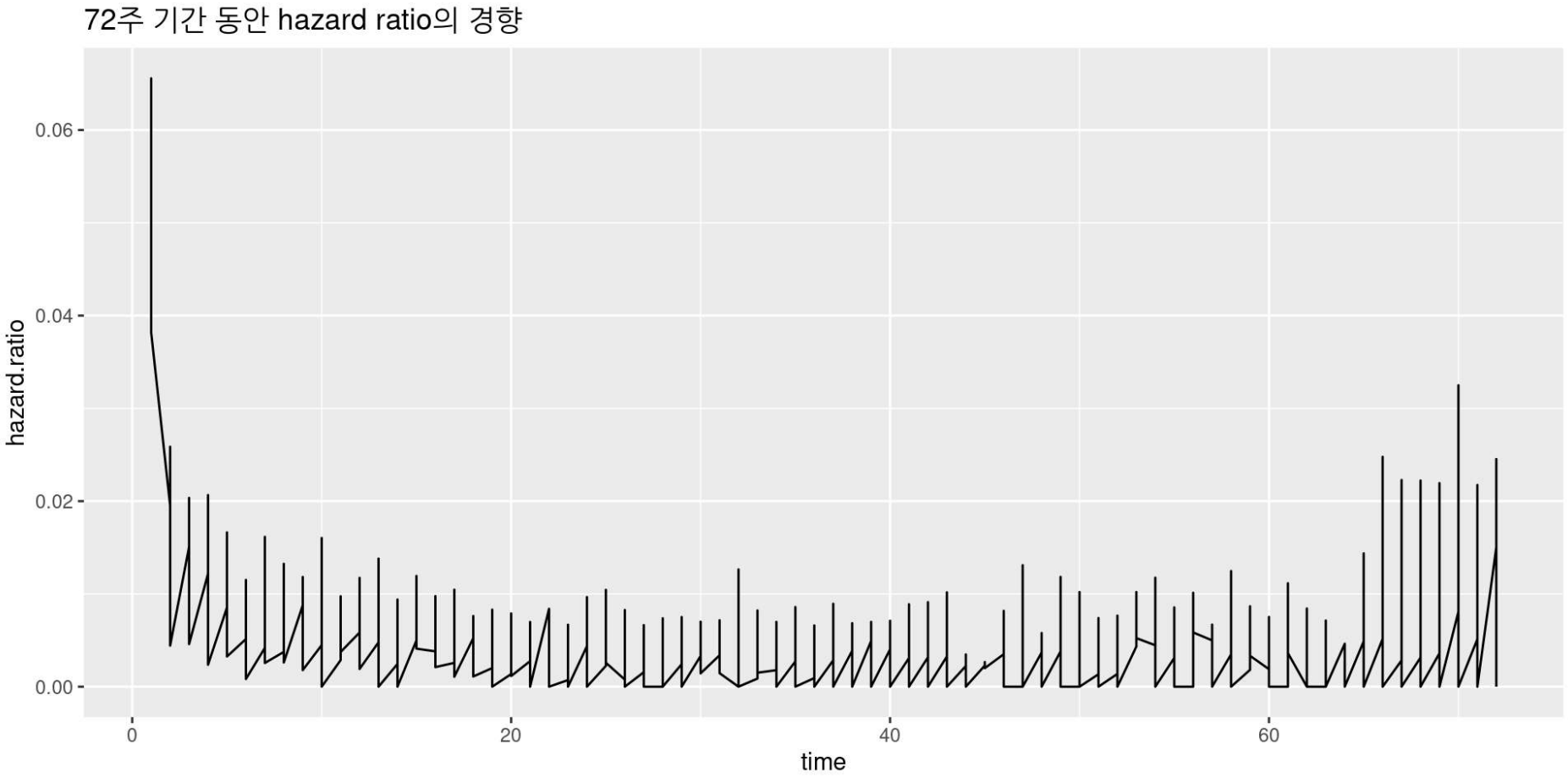
(Linoff and Berry 2011).

Kaplan-Meier 적합 시 생성되는 table을 활용하여 데이터로부터 경험적 Hazard ratio를 구할 수 있음.

구독 모델의 경우 초기 진입 이후 프로모션이 끝나면 이탈하는 고객이 있음으로 시간에 따른 Hazard의 모양은 U-자형으로 보임.

```
tibble(time=surv.fit$time,
       n.event=surv.fit$n.event,
       n.risk=surv.fit$n.risk) %>%
  mutate(hazard.ratio=n.event/n.risk) %>%
  ggplot() +
  geom_line(aes(x=time, y=hazard.ratio)) +
  ggtitle("72주 기간 동안 hazard ratio의 경향")
```

Hazard의 경향 확인



Cox 비례 위험 모형

```r
idx <- churnTel %>%
  mutate(idx = row_number()) %
  filter(internet.service != '
  filter(phone.service != "no'
  pull(idx)

data <- churnTel[idx, ] %>%
  mutate(across(where(is.facto
                droplevels)) %
  select(-c(customer.ID,
            phone.service,
            total.charge))

coxph.churn <-
  coxph(Surv(tenure, churn=="y
                     data=data
                     subset=tr
```

비례위험모형

$$\lambda_{\mathbf{x}_i}(t) = \lambda_0(t)e^{\beta^t \mathbf{x}_i}$$

Cox 비례 위험 모형을 구축함. phone, internet service를 모두 구독하는 고객에 대해서만 Cox 모델 적합을 수행함. (no phone or internet service 고객에 대해서도 모형 적합 시 일부 계수값이 정해지지 않음)

Cox 비례 위험 모형

```
1  coefficients(summary(coxph.churn))
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
gendermale	-0.085	0.92	0.06	-2e+00	1e-01
senior.citizen	-0.019	0.98	0.07	-3e-01	8e-01
partnerno	0.601	1.82	0.07	9e+00	4e-19
dependentno	0.022	1.02	0.09	3e-01	8e-01
multiple.linesno	0.526	1.69	0.15	4e+00	3e-04
internet.servicefiber-optic	-0.006	0.99	0.67	-9e-03	1e+00
online.securityno	0.811	2.25	0.16	5e+00	2e-07
online.backupno	0.799	2.22	0.15	5e+00	6e-08
device.protectionno	0.431	1.54	0.15	3e+00	4e-03
tech.supportno	0.461	1.59	0.15	3e+00	3e-03
streaming.TVno	0.170	1.18	0.27	6e-01	5e-01
streaming.movieno	0.250	1.28	0.27	9e-01	4e-01
contract1-year	-1.481	0.23	0.11	-1e+01	4e-44
contract2-year	-2.716	0.07	0.20	-1e+01	2e-43
paperless.billno	-0.200	0.82	0.07	-3e+00	4e-03
payment.methodcredit-card	-0.092	0.91	0.11	-8e-01	4e-01
payment.methodelectronic-check	0.555	1.74	0.08	7e+00	4e-11
payment.methodmail-check	0.486	1.63	0.11	4e+00	1e-05
monthly.charge	0.013	1.01	0.03	5e-01	6e-01

모형의 평가

$$c = Pr(y_i > y_j | x_i > x_j)$$

모형의 평가지표로 **Concordance index(c-index)**를 사용할 수 있음. 이는 모든 개체 짝 중 이벤트의 발생이 먼저 일어난 개체가 상대적 위험도 높은 짝의 비율을 계산한 것임(T. Therneau and Atkinson 2023). 수식은 위와 같음.

```
1 concordance(coxph.churn)
```

```
Call:
concordance.coxph(object = coxph.churn)

n= 3866
Concordance= 0.86 se= 0.0043
concordant discordant     tied.x     tied.y    tied.xy
   2819936     478008         24      39496          2
```

모형의 평가

Cox 비례 위험 모형의 `predict`는 아래와 같이 5가지의 type을 지원함(T. M. Therneau and Lumley 2015).

$$\hat{\lambda}_{\mathbf{x}_i}(t) = \hat{\lambda}_0(t) e^{\hat{\beta}^t \mathbf{x}_i}$$

- `lp`: the linear predictor $\hat{\beta}^t \mathbf{x}_i$
- `risk`: the risk score exp(lp) $e^{\hat{\beta}^t \mathbf{x}_i}$
- `terms`: the terms of the linear predictor $\hat{\beta}_1 \mathbf{x}_1, \cdots, \beta_p \mathbf{x}_p$
- `expected`: the expected number of events given the covariates and follow-up time $\int_0^t \hat{\lambda}_{\mathbf{x}_i}(t) dt$
- `survival`: The survival probability for a subject is equal to exp(-expected) $\hat{S}_{\mathbf{x}_i}(t) = -\exp(\int_0^t \hat{\lambda}_{\mathbf{x}_i}(t) dt)$

모형의 활용

- `lp`: the linear predictor $\hat{\beta}^t \mathbf{x}_i$

```
predict(coxph.churn, data[test, ][1:5, ], type = "lp", , reference="sample"
```

```
    5040     5748     3627     3750      880
 4.31965 -0.27185  2.36609  3.71940  2.31517
```

```
X <- subset(model.matrix(churn ~ ., data[test, ])[,2:21], select=-tenure)
new.X <- X - rep(coxph.churn$means, each=nrow(X))
(new.X %*% coef(coxph.churn))[1:5, ]
```

```
    5040     5748     3627     3750      880
 4.31965 -0.27185  2.36609  3.71940  2.31517
```

- `risk`: the risk score exp(lp) $e^{\hat{\beta}^t \mathbf{x}_i}$

```
predict(coxph.churn, data[test, ][1:5, ], type = "risk", reference="sample"
```

```
    5040     5748     3627     3750      880
75.16222  0.76196 10.65562 41.23981 10.12668
```

```
exp((new.X %*% coef(coxph.churn))[1:5, ])
```

```
    5040     5748     3627     3750      880
75.16222  0.76196 10.65562 41.23981 10.12668
```

모형의 활용

- `terms`: the terms of the linear predictor $\hat{\beta}_1 \mathbf{x}_1, \cdots, \beta_p \mathbf{x}_p$

```
predict(coxph.churn, data[test, ][1, ], type = "terms", reference="sample")
```

```
     gender senior.citizen partner dependent multiple.lines internet.service
5040      0      -0.019251  0.6011  0.021746        0.52554       -0.0062472
     online.security online.backup device.protection tech.support streaming.TV
5040         0.81088       0.79887           0.43078      0.46109      0.16965
     streaming.movie contract paperless.bill payment.method monthly.charge
5040               0        0              0        0.55515       -0.02966
attr(,"constant")
[1] 1.0357
```

```
coef(coxph.churn) * new.X[1, ]
```

모형의 활용

- `expected`: the expected number of events given the covariates and follow-up time
  $\int_0^t \hat{\lambda}_{\mathbf{x}_i}(t)dt$

```
predict(coxph.churn, data[test, ][1:5, ], type = "expected", reference="sar
```

```
[1] 0.396700 0.026959 0.271604 0.217660 0.368438
```

- `survival`: The survival probability for a subject is equal to exp(-expected)
  $\hat{S}_{\mathbf{x}_i}(t) = -\exp(\int_0^t \hat{\lambda}_{\mathbf{x}_i}(t)dt)$

```
predict(coxph.churn, data[test, ][1:5, ], type = "survival")
```

```
[1] 0.67254 0.97340 0.76216 0.80440 0.69181
```

```
exp(-predict(coxph.churn, data[test, ][1:5, ], type = "expected", reference
```

```
[1] 0.67254 0.97340 0.76216 0.80440 0.69181
```

모형의 활용

상대적 위험도의 값에 따라 고객을 10개의 그룹으로 분류한 후, 그룹 내에서 이탈 고객의 비중을 구하여 모델을 검증할 수 있음(Li 1995).

train 데이터셋의 결과

```
# A tibble: 10 × 4
# Groups:   groups [10]
   groups           yes    no churn.ratio
   <fct>          <int> <int>       <dbl>
 1 (-0.0381,8.87]   275  1648        14.3
 2 (8.87,17.7]      235   336        41.2
 3 (17.7,26.5]      215   211        50.5
 4 (26.5,35.3]      130   128        50.4
 5 (35.3,44.1]      108    86        55.7
 6 (44.1,52.9]      120    75        61.5
 7 (52.9,61.8]       64    41        61
 8 (61.8,70.6]       39    18        68.4
 9 (70.6,79.4]       42    24        63.6
10 (79.4,88.3]       52    19        73.2
```

test 데이터셋의 결과

```
# A tibble: 10 × 4
# Groups:   groups [10]
   groups           yes    no churn.ratio
   <fct>          <int> <int>       <dbl>
 1 (-0.0389,8.88]    71   410        14.8
 2 (8.88,17.7]       42    81        34.1
 3 (17.7,26.5]       38    58        39.6
 4 (26.5,35.4]       43    44        49.4
 5 (35.4,44.2]       34    25        57.6
 6 (44.2,53]         40    17        70.2
 7 (53,61.9]         10     9        52.6
 8 (61.9,70.7]        4     4        50
 9 (70.7,79.5]       11     5        68.8
10 (79.5,88.5]       13     7        65
```

참고문헌

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.

Li, Shaomin. 1995. “Survival Analysis.” *Marketing Research* 7 (4): 16.

Linoff, Gordon S, and Michael JA Berry. 2011. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. John Wiley & Sons.

Mohammadi, Reza, and Kevin Burke. 2023. *Liver: "Eating the Liver of Data Science"*. https://CRAN.R-project.org/package=liver.

Therneau, Terry M, and Thomas Lumley. 2015. “Package ‘Survival’.” *R Top Doc* 128 (10): 28–33.

Therneau, Terry, and Elizabeth Atkinson. 2023. “1 the Concordance Statistic.”

서영정. 2023. “머신러닝 기반 생존분석기법을 활용한 고객 이탈 예측 기술.” *Journal of Digital Contents Society* 24 (8): 1871–80.

허명회. 2023. “응용데이터분석방법론 4. 생존분석 강의노트.”