CaseStudy-03: Saving Customers

Nils Gandlau

14 11 2019

Data Preparation

```
str(dt)
```

```
## Classes 'data.table' and 'data.frame':   45017 obs. of  4 variables:
##  $ saveID    : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ offer     : chr  "E" "C" "E" "H" ...
##  $ saveMonth : num  5 5 6 3 1 2 3 3 5 5 ...
##  $ discoMonth: num  NA NA NA NA NA 3 NA NA NA 14 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

- Create a new variable `survivalTime` which we define as the difference `discoMonth - saveMonth`.
- Create a new variable `eventOccured` that takes on the value 1 if the customer discontinued the service within the 9-month observation period or 0 if he didn't.

```
dt[, saveID := NULL]
dt[, survivalTime := discoMonth - saveMonth]
dt[, eventOccurred := ifelse(is.na(discoMonth), 0, 1)]
```

Currently, for (right) censored observations we have `survivalTime == NA`. We will replace those `NA`s with the *total duration of the study*, which equals 9 month according to the case study handout.

```
dt[is.na(survivalTime), survivalTime := 9]
```

Looking at the description of the data set, we notice that offer "O" is the only case where no external incentive (e.g. coupon) was given to the customer. Instead, offer "O" represents the scenario where the customer simply "changed his mind" about canceling and retained with the firm on her own. **Hence, for all future models, we will use offer "O" as our reference group**, such that we can compare the effects of exogenous incentives as opposed to no exogenous incentives on survival time.

To make offer "O" the reference group, we will change the name such that our models pick it as reference group.

```
dt[offer == "O", offer := "_O"]
```

Cox Proportional Hazard Model

```
cox <- coxph(Surv(survivalTime, eventOccurred) ~ offer, data = dt)
summary(cox)
```

```
## Call:
## coxph(formula = Surv(survivalTime, eventOccurred) ~ offer, data = dt)
##
##   n= 45017, number of events= 16616
##
##            coef exp(coef) se(coef)       z Pr(>|z|)
## offerA -1.19937   0.30138  0.05921 -20.258  < 2e-16 ***
## offerB -1.08730   0.33712  0.03976 -27.350  < 2e-16 ***
## offerC -1.26859   0.28123  0.04891 -25.935  < 2e-16 ***
## offerD -1.45267   0.23394  0.06911 -21.020  < 2e-16 ***
## offerE -0.09307   0.91113  0.02815  -3.306 0.000946 ***
## offerF  0.04701   1.04813  0.02663   1.765 0.077548 .
## offerG -0.89178   0.40992  0.05875 -15.180  < 2e-16 ***
## offerH -0.55475   0.57422  0.04185 -13.257  < 2e-16 ***
## offerI -0.29336   0.74576  0.04208  -6.972 3.13e-12 ***
## offerJ -0.47215   0.62366  0.06178  -7.642 2.13e-14 ***
## offerK -0.35928   0.69818  0.03359 -10.694  < 2e-16 ***
## offerL -0.78190   0.45754  0.04968 -15.739  < 2e-16 ***
## offerM -0.96591   0.38064  0.05241 -18.431  < 2e-16 ***
## offerN -1.04414   0.35199  0.07791 -13.401  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## offerA    0.3014     3.3180    0.2684    0.3385
## offerB    0.3371     2.9663    0.3119    0.3644
## offerC    0.2812     3.5559    0.2555    0.3095
## offerD    0.2339     4.2745    0.2043    0.2679
## offerE    0.9111     1.0975    0.8622    0.9628
## offerF    1.0481     0.9541    0.9948    1.1043
## offerG    0.4099     2.4395    0.3653    0.4599
## offerH    0.5742     1.7415    0.5290    0.6233
## offerI    0.7458     1.3409    0.6867    0.8099
## offerJ    0.6237     1.6034    0.5525    0.7039
## offerK    0.6982     1.4323    0.6537    0.7457
## offerL    0.4575     2.1856    0.4151    0.5043
## offerM    0.3806     2.6272    0.3435    0.4218
## offerN    0.3520     2.8410    0.3021    0.4101
##
## Concordance= 0.629  (se = 0.002 )
## Likelihood ratio test= 3543  on 14 df,   p=<2e-16
## Wald test            = 3077  on 14 df,   p=<2e-16
## Score (logrank) test = 3369  on 14 df,   p=<2e-16
```

```
# Function for automated interpretation
interpretCoxPH <- function(coefName, expCoefValue){
  expCoefValue <- round(expCoefValue, 2)
  interpretation <- paste0(
    "At any time t, customers that received offer ",
    str_sub(coefName, -1),
    " have a risk that is  ",
    expCoefValue,
    " times as high as the reference offer O."
  )
}

# Create table that summarizes the results nicely
resultCox <- rbindlist(lapply(names(cox$coefficients), function(coefName){
  coefValue <- cox$coefficients[[coefName]]
  expCoefValue <- exp(coefValue)
  interpretation <- interpretCoxPH(coefName, expCoefValue)
  return(data.table(
    coefficient = coefName,
    value = coefValue,
    `exp(value)` = expCoefValue,
    interpretation = interpretation
  ))
}))

resultCox[, c("interpretation")]
```

interpretation

At any time t, customers that received offer A have a risk that is 0.3 times as high as the reference offer O.
At any time t, customers that received offer B have a risk that is 0.34 times as high as the reference offer O.
At any time t, customers that received offer C have a risk that is 0.28 times as high as the reference offer O.
At any time t, customers that received offer D have a risk that is 0.23 times as high as the reference offer O.
At any time t, customers that received offer E have a risk that is 0.91 times as high as the reference offer O.
At any time t, customers that received offer F have a risk that is 1.05 times as high as the reference offer O.
At any time t, customers that received offer G have a risk that is 0.41 times as high as the reference offer O.
At any time t, customers that received offer H have a risk that is 0.57 times as high as the reference offer O.
At any time t, customers that received offer I have a risk that is 0.75 times as high as the reference offer O.
At any time t, customers that received offer J have a risk that is 0.62 times as high as the reference offer O.
At any time t, customers that received offer K have a risk that is 0.7 times as high as the reference offer O.
At any time t, customers that received offer L have a risk that is 0.46 times as high as the reference offer O.
At any time t, customers that received offer M have a risk that is 0.38 times as high as the reference offer O.
At any time t, customers that received offer N have a risk that is 0.35 times as high as the reference offer O.

All coefficients are statistically significant.

In summary, these results indicate that customers who reveiced any of the offers except offer “F” are less likely to churn at any given period compared to those customers who did not receive an offer but simply changed their mind (refernce group offer “O”).

Offers “A” to “D” are very effective for improving longevity of the customer (relative to offer O).

Offer “D” seems to be the best offer among all. Customers in this group have almost half the risk of churning compared to those that received no incentive (offer “O”).

Accelerated failure time model (AFT)

In this section we fit an AFT model assuming a Weibull distribution for the hazard function.

```
aft <- survreg(Surv(survivalTime, eventOccurred) ~ offer,
               data = dt,
               dist="weibull")

summary(aft)
```

```
##
## Call:
## survreg(formula = Surv(survivalTime, eventOccurred) ~ offer,
##     data = dt, dist = "weibull")
##                Value Std. Error      z      p
## (Intercept)  2.40179    0.01321 181.79 <2e-16
## offerA       0.71271    0.03537  20.15 <2e-16
## offerB       0.64638    0.02391  27.03 <2e-16
## offerC       0.75398    0.02938  25.66 <2e-16
## offerD       0.86324    0.04132  20.89 <2e-16
## offerE       0.05172    0.01668   3.10 0.0019
## offerF      -0.03479    0.01577  -2.21 0.0274
## offerG       0.52963    0.03496  15.15 <2e-16
## offerH       0.32893    0.02487  13.23 <2e-16
## offerI       0.17019    0.02495   6.82  9e-12
## offerJ       0.28020    0.03664   7.65  2e-14
## offerK       0.21202    0.01994  10.63 <2e-16
## offerL       0.46506    0.02957  15.73 <2e-16
## offerM       0.57486    0.03125  18.39 <2e-16
## offerN       0.62096    0.04632  13.41 <2e-16
## Log(scale)  -0.52369    0.00724 -72.35 <2e-16
##
## Scale= 0.592
##
## Weibull distribution
## Loglik(model)= -63054.4   Loglik(intercept only)= -64866
##  Chisq= 3623.14 on 14 degrees of freedom, p= 0
## Number of Newton-Raphson Iterations: 5
## n= 45017
```

```
# Create a nicely formatted table that summarizes the AFT model's result
aftResult <- rbindlist(lapply(names(aft$coefficients), function(coefName){
  coefValue <- aft$coefficients[[coefName]]
  accelerationParameter <- round(exp(coefValue), 2)

  interpretation <- ""
  if (startsWith(coefName, "offer")){
    interpretation <- paste0(
      "A customer that received Offer ",
      str_sub(coefName, -1), # extract last character of string
      " lives ",
      accelerationParameter,
      " times as long as the reference group (offer O)."
    )
  }

  return(data.table(
    ` coefficient ` = coefName,
    ` value ` = round(coefValue, 4),
    `__exp(value)` = accelerationParameter,
    interpretation = interpretation
  ))
}))

aftResult[, 1:3]
```

coefficient	value	__exp(value)
(Intercept)	2.4018	11.04
offerA	0.7127	2.04
offerB	0.6464	1.91
offerC	0.7540	2.13
offerD	0.8632	2.37
offerE	0.0517	1.05
offerF	-0.0348	0.97
offerG	0.5296	1.70
offerH	0.3289	1.39
offerI	0.1702	1.19
offerJ	0.2802	1.32
offerK	0.2120	1.24
offerL	0.4651	1.59
offerM	0.5749	1.78
offerN	0.6210	1.86

```
aftResult[, c("interpretation")]
```

interpretation
A customer that received Offer A lives 2.04 times as long as the reference group (offer O).
A customer that received Offer B lives 1.91 times as long as the reference group (offer O).
A customer that received Offer C lives 2.13 times as long as the reference group (offer O).
A customer that received Offer D lives 2.37 times as long as the reference group (offer O).
A customer that received Offer E lives 1.05 times as long as the reference group (offer O).
A customer that received Offer F lives 0.97 times as long as the reference group (offer O).
A customer that received Offer G lives 1.7 times as long as the reference group (offer O).

interpretation

A customer that received Offer H lives 1.39 times as long as the reference group (offer O).
A customer that received Offer I lives 1.19 times as long as the reference group (offer O).
A customer that received Offer J lives 1.32 times as long as the reference group (offer O).
A customer that received Offer K lives 1.24 times as long as the reference group (offer O).
A customer that received Offer L lives 1.59 times as long as the reference group (offer O).
A customer that received Offer M lives 1.78 times as long as the reference group (offer O).
A customer that received Offer N lives 1.86 times as long as the reference group (offer O).

All coefficients are statistically significant.

The results are in line with those of the Cox Proportional Hazard Model.

In summary, the AFT model shows that customers who received any of the offers except offer F live on average longer than a customer that decided to remain with the firm on his own (ie, reference group offer “O”).

Offers “A” to “D” are very effective improving longevity of the customer (relative to offer O).

Offer “D” seems to be the best offer among all.

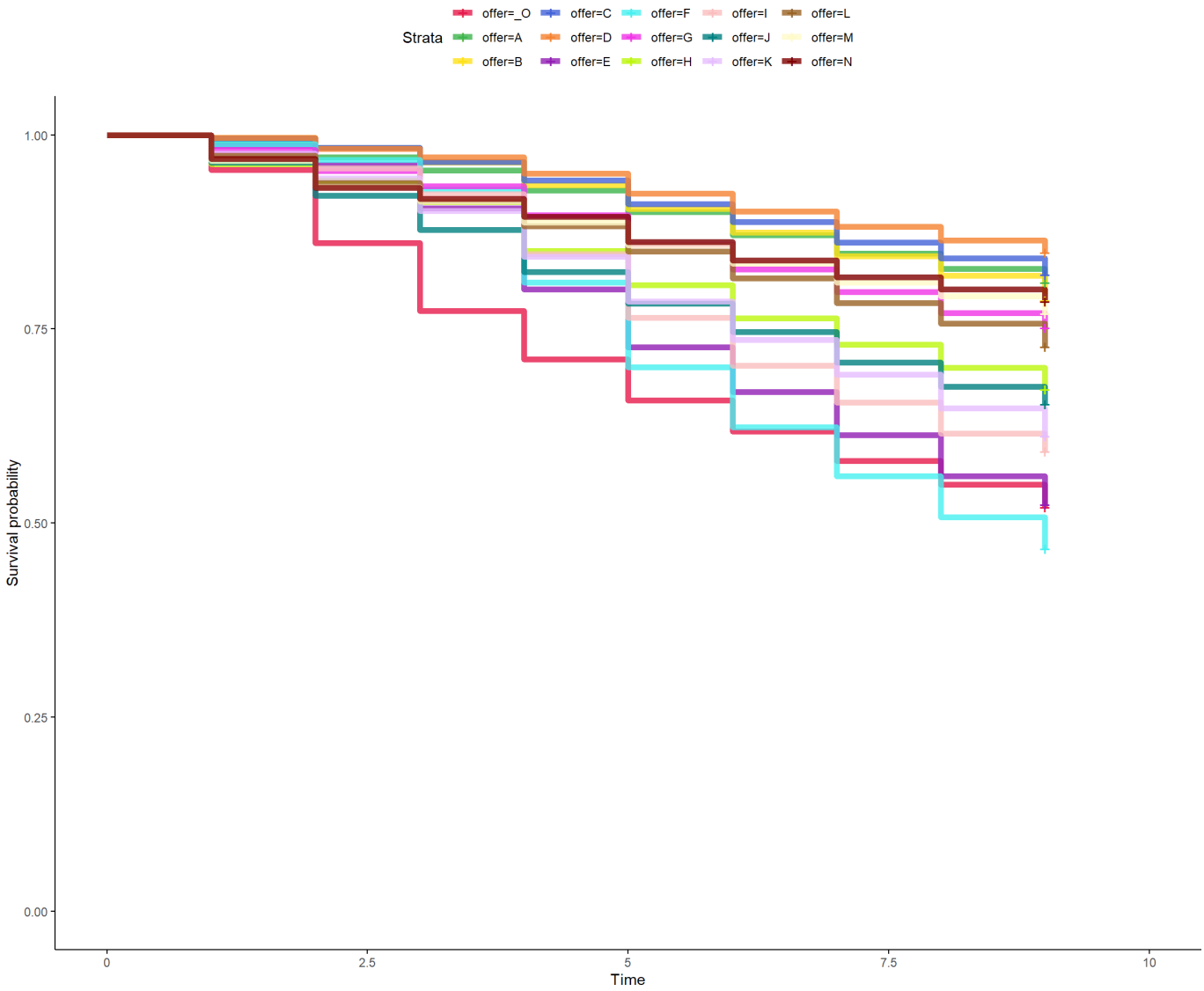
Kaplan-Meier

Finally, we fit a non-parametric Kaplan-Meier model to the data. We then visualize the different survival functions.

The results are in line with both the (1) Cox Proportional Hazard Model and (2) the AFT model.

```
km <- survfit(Surv(survivalTime, eventOccurred) ~ offer,
              data = dt,
              type = "kaplan-meier")
```

```
ggsurvplot(km,
           ggtheme = theme_classic(),
           size=2,
           alpha = 0.8,
           palette = c('#e6194b', '#3cb44b', '#ffe119', '#4363d8', '#f58231', '#911eb4', '#46
f0f0', '#f032e6', '#bcf60c', '#fabebe', '#008080', '#e6beff', '#9a6324', '#fffac8', '#800000'
, '#aaffc3', '#808000', '#ffd8b1', '#000075', '#808080', '#ffffff', '#000000'))
```



- Looking closely, one can observe that the survival functions of offers A, B, C, D are almost always above any other survival function. This reinforces the results that we have gotten from the other models, namely that offers A-D are affecting longevity of customers most positively.
- The baseline survival function, resembling the survival function of customers who received offer D has the lowest survival rates at any given period except period 9.
- In period 9, amongst others offers J and N have a less (or equal) survival probability than customers who received offer D

The firm should use incentives, but ideally incentives of types A to D, since they affect longevity most positively.