PA Final Project

Group 7

2023-11-29

Import Data

```
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
```

```
np = read.table("np.csv", header=T, na.strings=".") %>%
                arrange(SubscriptionId, t) %>%
                group_by(SubscriptionId) %>%
                mutate(nextchurn = lead(churn), nextprice=lead(currprice), t = as.factor(t))
```

(a)

```
table(np$churn)
```

```
##
##     0     1
## 11681   489
```

```
summary(np)
```

```
##  SubscriptionId           t            churn           regularity
##  Length:12170       1      :2064   Min.   :0.00000   Min.   : 0.000
##  Class :character   2      :1846   1st Qu.:0.00000   1st Qu.: 0.000
##  Mode  :character   3      :1663   Median :0.00000   Median : 4.000
##                     4      :1503   Mean   :0.04018   Mean   : 8.106
```

```
##                   5      :1319   3rd Qu.:0.00000   3rd Qu.:14.000
##                   6      :1151   Max.   :1.00000   Max.   :30.000
##                   (Other):2624
##    intensity          sports1            news1            crime1
##  Min.   :  0.000   Min.   :   0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  7.433   Median :   1.00   Median :  1.00   Median :  0.000
##  Mean   : 10.765   Mean   :  12.38   Mean   :  9.58   Mean   :  5.094
##  3rd Qu.: 14.798   3rd Qu.:  11.00   3rd Qu.:  9.00   3rd Qu.:  5.000
##  Max.   :332.435   Max.   :1129.00   Max.   :455.00   Max.   :176.000
##
##      life1             obits1         business1         opinion1
##  Min.   :  0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000
##  1st Qu.:  0.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.0000
##  Median :  0.000   Median : 0.000   Median : 0.000   Median : 0.0000
##  Mean   :  3.803   Mean   : 1.128   Mean   : 1.396   Mean   : 0.6053
##  3rd Qu.:  3.000   3rd Qu.: 0.000   3rd Qu.: 1.000   3rd Qu.: 0.0000
##  Max.   :369.000   Max.   :95.000   Max.   :71.000   Max.   :63.0000
##
##      mobile            tablet           desktop            loc1
##  Min.   :  0.000   Min.   :  0.000   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.00
##  Median :  0.000   Median :  0.000   Median :  0.00   Median :  0.00
##  Mean   :  7.586   Mean   :  3.495   Mean   : 14.89   Mean   : 12.78
##  3rd Qu.:  1.000   3rd Qu.:  0.000   3rd Qu.: 16.00   3rd Qu.:  9.00
##  Max.   :289.000   Max.   :260.000   Max.   :593.00   Max.   :514.00
##
##       Loc2              Loc3              Loc4           SrcGoogle
##  Min.   :  0.000   Min.   :  0.000   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.000   1st Qu.:  0.000
##  Median :  0.000   Median :  0.000   Median :  0.000   Median :  0.000
##  Mean   :  2.023   Mean   :  1.892   Mean   :  2.122   Mean   :  8.854
##  3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.:  0.000   3rd Qu.:  4.000
##  Max.   :188.000   Max.   :231.000   Max.   :235.000   Max.   :288.000
##
##    SrcDirect          SrcElm         SrcSocial       SrcBingYahooAol
##  Min.   :  0.000   Min.   :  0.00   Min.   :  0.000   Min.   :  0.000
##  1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.000   1st Qu.:  0.000
##  Median :  0.000   Median :  0.00   Median :  0.000   Median :  0.000
##  Mean   :  2.376   Mean   :  0.44   Mean   :  2.692   Mean   :  1.509
##  3rd Qu.:  0.000   3rd Qu.:  0.00   3rd Qu.:  0.000   3rd Qu.:  0.000
##  Max.   :250.000   Max.   :102.00   Max.   :182.000   Max.   :330.000
##
##  SrcNewsletter       SrcLegacy        SrcGoogleNews      SrcGoogleAd
##  Min.   :  0.000   Min.   :  0.0000   Min.   : 0.0000   Min.   :  0.0000
##  1st Qu.:  0.000   1st Qu.:  0.0000   1st Qu.: 0.0000   1st Qu.:  0.0000
##  Median :  0.000   Median :  0.0000   Median : 0.0000   Median :  0.0000
##  Mean   :  2.145   Mean   :  0.7949   Mean   : 0.1748   Mean   :  0.3482
##  3rd Qu.:  0.000   3rd Qu.:  0.0000   3rd Qu.: 0.0000   3rd Qu.:  0.0000
##  Max.   :137.000   Max.   :212.0000   Max.   :99.0000   Max.   :170.0000
##
##    currprice          trial           nextchurn        nextprice
##  Min.   : 0.00   Min.   :0.0000   Min.   :0.0000   Min.   : 3.33
##  1st Qu.:12.99   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:17.29
```

```
##  Median :19.99   Median :0.0000   Median :0.0000   Median :19.99
##  Mean   :15.85   Mean   :0.1149   Mean   :0.0401   Mean   :17.90
##  3rd Qu.:19.99   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:19.99
##  Max.   :19.99   Max.   :1.0000   Max.   :1.0000   Max.   :19.99
##  NA's   :689                      NA's   :2064     NA's   :2635
```

(b) - only consider the effects of trial, price, regularity and intensity

Analysis

```
np$t <- as.numeric(levels(np$t))[np$t]

# Run logistic regression models
model1 <- glm(nextchurn ~ t + trial + nextprice + regularity + intensity, np, family = binomial)
model2 <- glm(nextchurn ~ t + trial + nextprice + regularity, np, family = binomial)
model3 <- glm(nextchurn ~ t + trial + nextprice + intensity, np, family = binomial)

# Display model summaries
summary(model1)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + regularity +
##     intensity, family = binomial, data = np)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.002903   0.353028 -11.339  < 2e-16 ***
## t           -0.143106   0.029130  -4.913 8.98e-07 ***
## trial        0.360129   0.155889   2.310 0.020879 *
## nextprice    0.087507   0.018557   4.716 2.41e-06 ***
## regularity  -0.026510   0.007067  -3.751 0.000176 ***
## intensity   -0.007711   0.005163  -1.494 0.135285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3131.5  on 9529  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3143.5
##
## Number of Fisher Scoring iterations: 6
```

```
summary(model2)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + regularity,
##     family = binomial, data = np)
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.049385   0.351291 -11.527  < 2e-16 ***
## t           -0.139531   0.028928  -4.823 1.41e-06 ***
## trial        0.346632   0.155260   2.233   0.0256 *
## nextprice    0.087371   0.018532   4.715 2.42e-06 ***
## regularity  -0.031944   0.006153  -5.192 2.08e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3134.0  on 9530  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3144
##
## Number of Fisher Scoring iterations: 6
```

```
summary(model3)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + intensity,
##     family = binomial, data = np)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.994132   0.351288 -11.370  < 2e-16 ***
## t           -0.130642   0.029002  -4.505 6.65e-06 ***
## trial        0.325119   0.155468   2.091 0.036507 *
## nextprice    0.079342   0.018338   4.327 1.51e-05 ***
## intensity   -0.018857   0.005002  -3.770 0.000163 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3146.0  on 9530  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3156
##
## Number of Fisher Scoring iterations: 6
```

Model summaries shown above.

```
variables_of_interest <- c("t", "trial", "nextprice", "regularity", "intensity")

# Create a new data frame with numeric columns
np_numeric <- np[, variables_of_interest]
np_numeric[variables_of_interest] <- lapply(np[variables_of_interest], as.numeric)
```

```r
# Correlation matrix
cor_matrix <- cor(select(np_numeric, all_of(variables_of_interest)), use = "complete.obs")
print("Correlation Matrix:")
```

```
## [1] "Correlation Matrix:"
```

```r
print(cor_matrix)
```

```
##                     t       trial   nextprice regularity   intensity
## t           1.0000000 -0.47865664  0.13873021 -0.2576637 -0.19693196
## trial      -0.4786566  1.00000000 -0.05506894  0.1844736  0.14216258
## nextprice   0.1387302 -0.05506894  1.00000000  0.1012444  0.03542222
## regularity -0.2576637  0.18447356  0.10124444  1.0000000  0.48545901
## intensity  -0.1969320  0.14216258  0.03542222  0.4854590  1.00000000
```

Negative Correlation between t and trial: There is a negative correlation (-0.48) between the month variable (t) and the trial variable. This indicates that, on average, as the month number increases, the likelihood of having a trial decreases. This negative correlation suggests that trial offers are more common in the earlier months.

Positive Correlation between trial and regularity and intensity: There is a positive correlation between the trial variable and both regularity (0.18) and intensity (0.14). This suggests that, on average, customers who have trial offers might also have higher engagement (more regularity and intensity) during the current month.

Weak Positive Correlation between nextprice and regularity and intensity: There is a weak positive correlation between the next price (nextprice) and both regularity (0.10) and intensity (0.04). This suggests that higher next prices are weakly associated with higher engagement.

Negative Correlation between t and regularity and intensity: There are negative correlations between the month variable (t) and both regularity (-0.26) and intensity (-0.20). This suggests that, on average, as the month number increases, there is a decrease in both regularity and intensity.

Moderate correlation between regularity and intensity: There is a positive correlation (0.49) between regularity and intensity. This is expected, as higher regularity (more reading days) is likely associated with higher intensity (more page views per reading day).

```r
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

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

```r
vif(model1)
```

```
##          t      trial  nextprice regularity  intensity
##   1.495581   1.449884   1.035239   1.463987   1.432546
```

```
vif(model2)
```

```
##          t      trial  nextprice regularity
##   1.484643   1.438270   1.035160   1.101866
```

```
vif(model3)
```

```
##         t     trial nextprice intensity
##  1.485314  1.445492  1.022144  1.095508
```

All VIF values are well below the commonly used threshold of 5, indicating that there is no severe multicollinearity in any of the models. => But why `intensity` is not significant?

The VIF values suggest that the predictors in each model are not highly correlated, and there is no evidence of problematic collinearity.

Answer:

What do you conclude about the effects of trial, price, regularity and intensity?

(a) What is the trial effect telling you, given that most trial offers are 1 month, many customers did not have trial offers, and you already have a dummy for month 1 in the model with the t variable?

the trial variable is positively associated with the likelihood of churn (~0.3), indicating that customers with trial offers are more likely to churn.

****But how about defining the length of trial offers, and what does it mean by 'already having a dummy for month 1 in the model with the t variable'?***

Customers who have trial offers are more likely to churn compared to those who did not have trial offers. This may seem counterintuitive at first, considering that trial offers are typically used to attract and retain customers. However, it's important to note that the trial effect could be capturing the impact of factors associated with the trial, such as the possibility that customers on trial offers may be exploring the service but decide to churn after the trial period.

(b) What do you conclude about the effects of intensity versus regularity? Which one should an organization develop strategies to encourage?

Model 1: Intensity not significant, due to the multicolinarity between intensity and regularity

Model 2 & 3: Inform us that effect: regularity » intensity and we should only include regularity in the model.

The `regularity` variable has a negative coefficient, and its statistical significance suggests that a higher number of reading days is associated with a lower likelihood of churn. Given that regularity is a significant predictor and has a negative association with churn, consider developing strategies to encourage regular engagement among users, that is, increase the number of reading days for each month.

i.e. What is reading days???

Moreover,

the nextprice variable suggests that an increase in the price paid next month (~0.07) is associated with higher likelihood of churn.

In summary, the organization may benefit from focusing on strategies that promote regular engagement among users, as this appears to be associated with a lower likelihood of churn. Also, offering a trial period of 1 month seems to be having a positive effect on the customers, so the company should also monitor customer behavior around that and provide this period to more customers.

(c) - adding the content variables and test `regularity` variable by control experiments

```
model_without_regularity <- glm(nextchurn ~ t + trial + nextprice + sports1 +
                                news1 + crime1 + life1 + obits1 + business1 + opinion1,
                                np, family = binomial)
summary(model_without_regularity)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + sports1 + news1 +
##     crime1 + life1 + obits1 + business1 + opinion1, family = binomial,
##     data = np)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.117635   0.350838 -11.737  < 2e-16 ***
## t           -0.130609   0.028793  -4.536 5.73e-06 ***
## trial        0.309241   0.154966   1.996   0.0460 *
## nextprice    0.083194   0.018471   4.504 6.67e-06 ***
## sports1     -0.006065   0.002528  -2.399   0.0164 *
## news1       -0.012748   0.005946  -2.144   0.0320 *
## crime1       0.008753   0.007843   1.116   0.2644
## life1        0.003819   0.008398   0.455   0.6493
## obits1      -0.009301   0.013787  -0.675   0.4999
## business1   -0.013241   0.026799  -0.494   0.6213
## opinion1     0.026258   0.027764   0.946   0.3443
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3140.3  on 9524  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3162.3
##
## Number of Fisher Scoring iterations: 6
```

Content Variables: `sports1` and `news1` are significant.

Trial and Price: `trial` and `nextprice` are still significant.

Other Content Variables: `crime1`, `life1`, `obits1`, `business1`, and `opinion1` are not significant.

```
model_with_regularity <- glm(nextchurn ~ t + trial + nextprice + sports1 + news1
                             + crime1 + life1 + obits1 + business1 + opinion1 + regularity,
                             np, family = binomial)
summary(model_with_regularity)
```

```
##
## Call:
```

```
## glm(formula = nextchurn ~ t + trial + nextprice + sports1 + news1 +
##     crime1 + life1 + obits1 + business1 + opinion1 + regularity,
##     family = binomial, data = np)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.0627006  0.3528121 -11.515  < 2e-16 ***
## t           -0.1420869  0.0290650  -4.889 1.02e-06 ***
## trial        0.3389870  0.1555876   2.179  0.02935 *
## nextprice    0.0882696  0.0186122   4.743 2.11e-06 ***
## sports1     -0.0006959  0.0027814  -0.250  0.80243
## news1       -0.0087485  0.0057577  -1.519  0.12865
## crime1       0.0108529  0.0076083   1.426  0.15373
## life1        0.0046254  0.0079180   0.584  0.55911
## obits1      -0.0012448  0.0137040  -0.091  0.92762
## business1   -0.0094182  0.0265700  -0.354  0.72299
## opinion1     0.0216623  0.0268220   0.808  0.41930
## regularity  -0.0288405  0.0090123  -3.200  0.00137 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3130.4  on 9523  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3154.4
##
## Number of Fisher Scoring iterations: 6
```

Now, all the content variables are not significant.

The `regularity` is significant and its sign is negative. This suggests that, holding other variables constant, an increase in regular reading days is associated with a decrease in the log-odds of churn.

AIC for the model with regularity is 3154.4, while without it was 3162.3. The lower AIC suggests that the model with regularity is a better fit for the data in terms of balancing goodness of fit and simplicity.

(d) - adding the device variables and test `regularity` variable by control experiments

```
model_device <- glm(nextchurn ~ t + trial + nextprice + mobile + tablet + desktop,
                    np, family = binomial)
summary(model_device)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + mobile + tablet +
##     desktop, family = binomial, data = np)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -4.086168   0.350129 -11.670  < 2e-16 ***
## t           -0.129269   0.028701  -4.504 6.67e-06 ***
## trial        0.307087   0.154737   1.985   0.0472 *
## nextprice    0.083129   0.018413   4.515 6.34e-06 ***
## mobile      -0.003066   0.002241  -1.368   0.1712
## tablet      -0.007853   0.004065  -1.932   0.0534 .
## desktop     -0.009112   0.002288  -3.983 6.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3138.5  on 9528  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3152.5
##
## Number of Fisher Scoring iterations: 6
```

Device Variables: `desktop` is significant.

Trial and Price: `trial` and `nextprice` are still significant.

Other Device Variables: `mobile` and `tablet` are not significant.

```
model_device_with_regularity <- glm(nextchurn ~ t + trial + nextprice + mobile + tablet + desktop
                                    + regularity, np, family = binomial)
summary(model_device_with_regularity)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + mobile + tablet +
##     desktop + regularity, family = binomial, data = np)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.056118   0.351543 -11.538  < 2e-16 ***
## t           -0.139194   0.028998  -4.800 1.59e-06 ***
## trial        0.339821   0.155496   2.185  0.02886 *
## nextprice    0.087613   0.018553   4.722 2.33e-06 ***
## mobile       0.001704   0.002761   0.617  0.53708
## tablet      -0.001022   0.004529  -0.226  0.82147
## desktop     -0.002576   0.003062  -0.841  0.40019
## regularity  -0.028474   0.010969  -2.596  0.00943 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3131.9  on 9527  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3147.9
##
## Number of Fisher Scoring iterations: 6
```

Now, all the device variables are not significant.

The `regularity` is significant and its sign is negative. This suggests that, holding other variables constant, an increase in regular reading days is associated with a decrease in the log-odds of churn.

AIC for the model with regularity is 3358.2, while without it was 3152.5. The higher AIC suggests that the model with regularity is a worse fit for the data in terms of balancing goodness of fit and simplicity.

So we conclude that those device variables will be permenantly dismissed by comparing the answers from part (c) and (d).

(e)

```
model_all <- glm(nextchurn ~ t + trial + nextprice + sports1 + news1
                             + crime1 + life1 + obits1 + business1 + opinion1
                             + mobile + tablet + desktop,  np, family = binomial)
summary(model_all)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + sports1 + news1 +
##     crime1 + life1 + obits1 + business1 + opinion1 + mobile +
##     tablet + desktop, family = binomial, data = np)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.110599   0.351710 -11.687  < 2e-16 ***
## t           -0.133747   0.028873  -4.632 3.62e-06 ***
## trial        0.309941   0.155198   1.997   0.0458 *
## nextprice    0.085419   0.018532   4.609 4.04e-06 ***
## sports1     -0.002895   0.003002  -0.964   0.3348
## news1       -0.009221   0.006105  -1.510   0.1310
## crime1       0.009176   0.007918   1.159   0.2465
## life1        0.006672   0.008501   0.785   0.4326
## obits1      -0.003946   0.014037  -0.281   0.7786
## business1   -0.006799   0.026958  -0.252   0.8009
## opinion1     0.024690   0.028037   0.881   0.3785
## mobile      -0.001506   0.003009  -0.501   0.6167
## tablet      -0.005639   0.004848  -1.163   0.2448
## desktop     -0.007205   0.003121  -2.308   0.0210 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3133.8  on 9521  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3161.8
##
## Number of Fisher Scoring iterations: 6
```

```
model_all_with_regularity <- glm(nextchurn ~ t + trial + nextprice + sports1 + news1
                                 + crime1 + life1 + obits1 + business1 + opinion1 + mobile +
                                   tablet + desktop + regularity, np, family = binomial)
summary(model_all_with_regularity)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + sports1 + news1 +
##     crime1 + life1 + obits1 + business1 + opinion1 + mobile +
##     tablet + desktop + regularity, family = binomial, data = np)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.0696842  0.3530572 -11.527  < 2e-16 ***
## t           -0.1419310  0.0291129  -4.875 1.09e-06 ***
## trial        0.3353856  0.1557784   2.153   0.0313 *
## nextprice    0.0886031  0.0186249   4.757 1.96e-06 ***
## sports1     -0.0007571  0.0029725  -0.255   0.7989
## news1       -0.0088498  0.0059351  -1.491   0.1359
## crime1       0.0097329  0.0077780   1.251   0.2108
## life1        0.0053028  0.0082617   0.642   0.5210
## obits1       0.0005644  0.0137995   0.041   0.9674
## business1   -0.0066273  0.0266803  -0.248   0.8038
## opinion1     0.0232581  0.0273815   0.849   0.3957
## mobile       0.0021894  0.0032559   0.672   0.5013
## tablet      -0.0008314  0.0049724  -0.167   0.8672
## desktop     -0.0019704  0.0036092  -0.546   0.5851
## regularity  -0.0273081  0.0116753  -2.339   0.0193 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3128.5  on 9520  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3158.5
##
## Number of Fisher Scoring iterations: 6
```

Same conclusions if fitting a model with payment, content, and device variables all in at the same time.

```
library(glmnet)
```

```
## Loading required package: Matrix
```

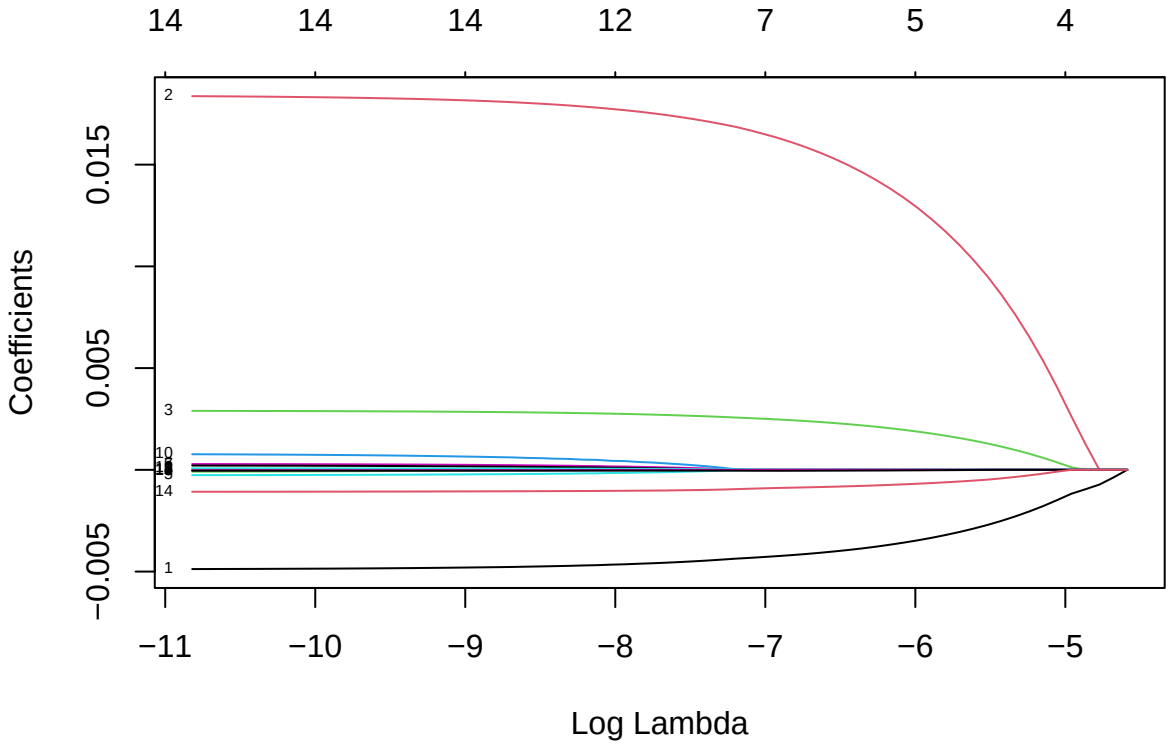
```
## Loaded glmnet 4.1-8
```

```
X <- np[, c('t', 'trial', 'nextprice', 'sports1', 'news1', 'crime1', 'life1', 'obits1', 'business1', 'op
Y <- np$nextchurn
```

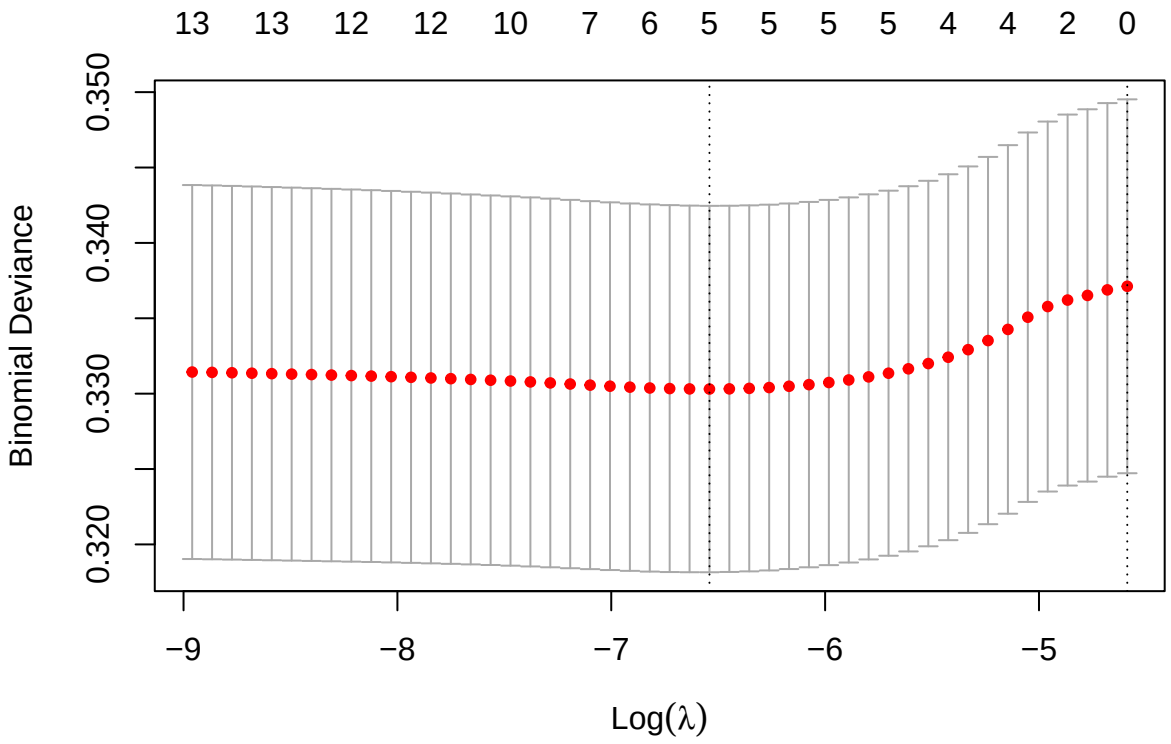
```r
complete_cases <- complete.cases(X)
X_clean <- X[complete_cases, ]
Y_clean <- Y[complete_cases]

X_matrix <- as.matrix(X_clean)

# Use cross-validated Lasso logistic regression
cv_lasso <- cv.glmnet(X_matrix, Y_clean, family = "binomial", alpha = 1)
plot(glmnet(X_matrix, Y_clean, alpha=1), xvar='lambda', label='T')
```




```r
plot(cv_lasso)
```




```r
# Select the best lambda value
best_lambda <- cv_lasso$lambda.min
print(best_lambda)
```

```
## [1] 0.001443371
```

```r
# Fit the final model using the selected lambda
lasso_model <- glmnet(X_matrix, Y_clean, family = "binomial", alpha = 1, lambda = best_lambda)
```

Also got the same conclusion.

(f)

Factors `t` and `regularity` retain customers.

Factors `trial` and `nextprice` drive customers away. ???

`Content` and `Device` variables have no substantial effect on churn. We could also dismiss the `intensity` factor because of the existence of `regularity` factor.