Final_Project

Hongkai Lou

2024-11-24

```r
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
```

```r
library(car)
```

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

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

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

```r
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.95 loaded
```

```r
library(glmnet)
```

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

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

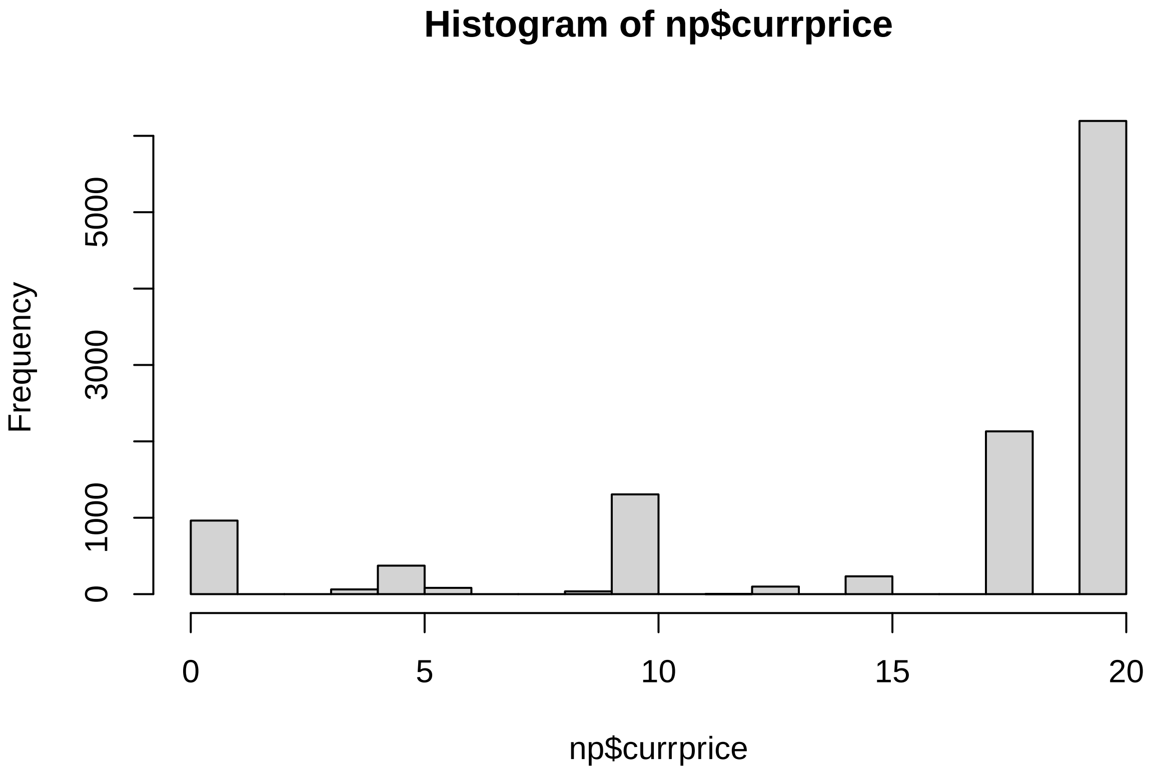
```r
np = read.table("np.csv", header=T, na.strings=".") %>%
  arrange(SubscriptionId, t) %>%
  group_by(SubscriptionId) %>%
  mutate(nextchurn = lead(churn),
  nextprice=lead(currprice),
  t = t)
table(np$nextchurn)
```

```
##
##    0    1
## 9701  405
```

```
table(np$trial)
```

```
##
##     0     1
## 10772  1398
```

```
hist(np$currprice)
```



2a

The trial has an positive effect in the model, meaning that it positively increase the churn possibility of a customer. Customers who are currently on a trial offer are more likely to churn. The reason might be that the trials are only offered to new customers (who registered the first time), and these users are not fully committed to the service and may view this trial as an opportunity to explore without an intention to continue.

2b

We see that regularity has a more significant impact on churn possibility of customers.

- First, notice that both regularity and intensity has a negative coefficient, meaning that a higher engagement overall will reduce the churn rate.

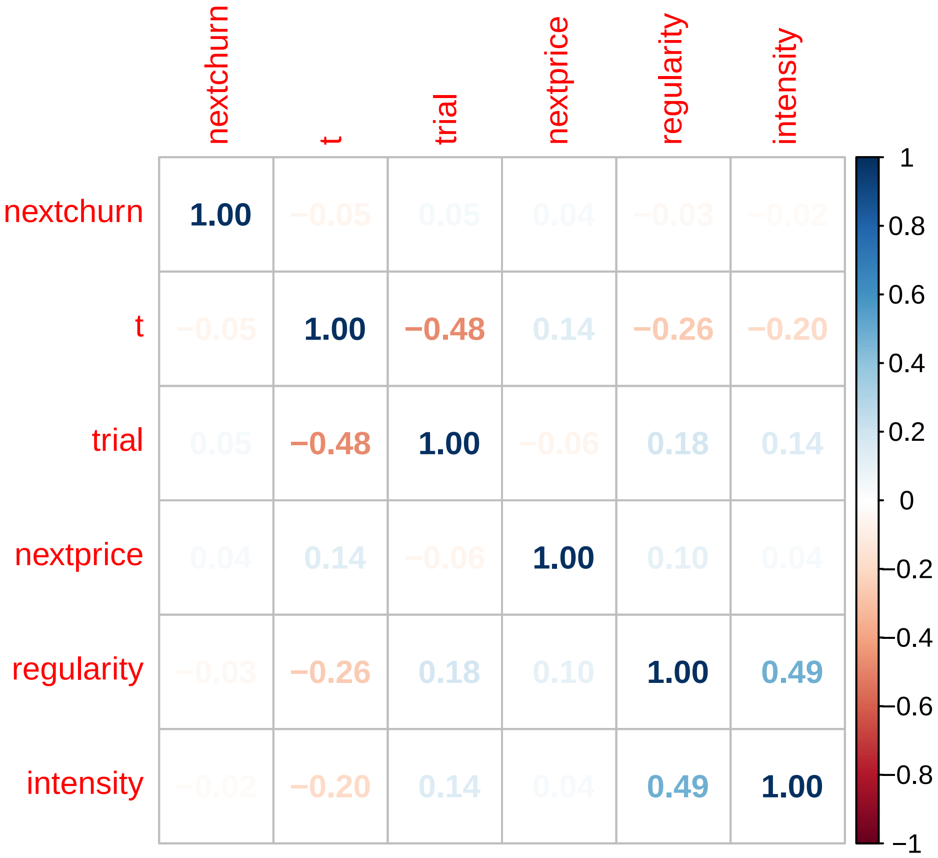
- Secondly, from the first model, where regularity and intensity are both included, we see that regularity has a higher t-value, meaning it is more statistically significant when they are grouped together. Furthermore, by displaying the vif, we observe that there is barely a multicollinearity situation here. Thus, as they are not highly associated, regularity is a more effective in interpreting churn rate than intensity when putting together

- Furthermore, by looking at the model separartely (lm2 and lm3), we see that regularity also has a smaller p-value. The residual deviance is also lower for the model including regularity. Thus, regularity is more effect, and the organization should encourage regularity. The reason why regularity is more important might due to the fact that higher intensity may cause readers to be 'burned out' in a short period of time, and may not be encouraged.

• In terms of nextprice, we see that nextprice variable is highly statistically significant in all three models, with a positive coefficient, which means that a higher price in next month will lead to a higher churn rate. That is a highly expected result, as customers will be less motivated to spend more money on the product.

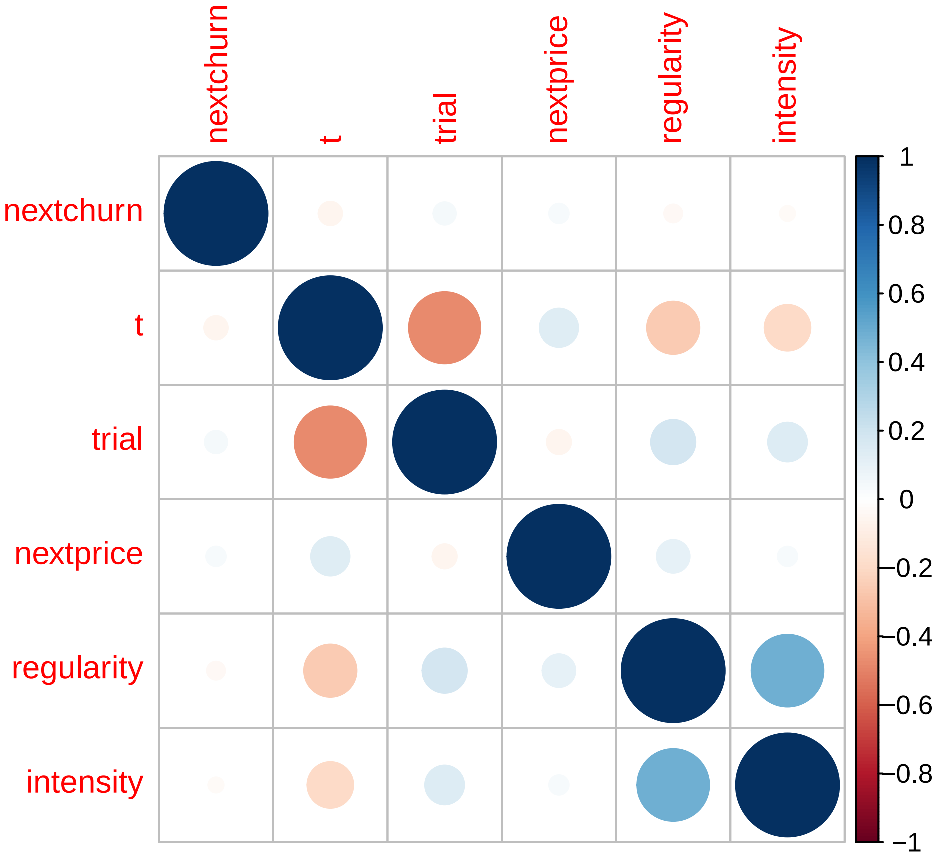
• In the end, all three models have small vif values, all variables have vif values less than 1.5, meaning there is little multicollinearity situation in the current analysis.

```
m = cor(np[,c('nextchurn','t', "trial", "nextprice", "regularity", "intensity")], use = 'complete.obs')
corrplot(m, method = 'number')
```



	nextchurn	t	trial	nextprice	regularity	intensity
nextchurn	1.00	−0.05	0.05	0.04	−0.03	−0.02
t	−0.05	1.00	−0.48	0.14	−0.26	−0.20
trial	0.05	−0.48	1.00	−0.06	0.18	0.14
nextprice	0.04	0.14	−0.06	1.00	0.10	0.04
regularity	−0.03	−0.26	0.18	0.10	1.00	0.49
intensity	−0.02	−0.20	0.14	0.04	0.49	1.00

```
corrplot(m)
```



```
np = na.omit(np)
lm1 <- glm(nextchurn ~ t+trial+nextprice+regularity+intensity, family = 'binomial', data = np)
summary(lm1)
```

```
##
## 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
## AIC: 3143.5
##
## Number of Fisher Scoring iterations: 6
```

```r
vif(lm1)
```

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

```r
lm2 <- glm(nextchurn ~ t+trial+nextprice+regularity, family = 'binomial', data = np)
summary(lm2)
```

```
##
## 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
## AIC: 3144
##
## Number of Fisher Scoring iterations: 6
```

```r
vif(lm2)
```

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

```r
lm3 <- glm(nextchurn ~ t+trial+nextprice+intensity, family = 'binomial',data = np)
summary(lm3)
```

```
##
## 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
## AIC: 3156
##
## Number of Fisher Scoring iterations: 6
```

```
vif(lm3)
```

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

3

• Content variables (e.g., sports, news, crime, life, obits, business, opinion) have varying effects on churn rate, but most of them show limited statistical significance. While some content types slightly influence the churn rate, their effects are generally weaker compared to factors like trial and nextprice.

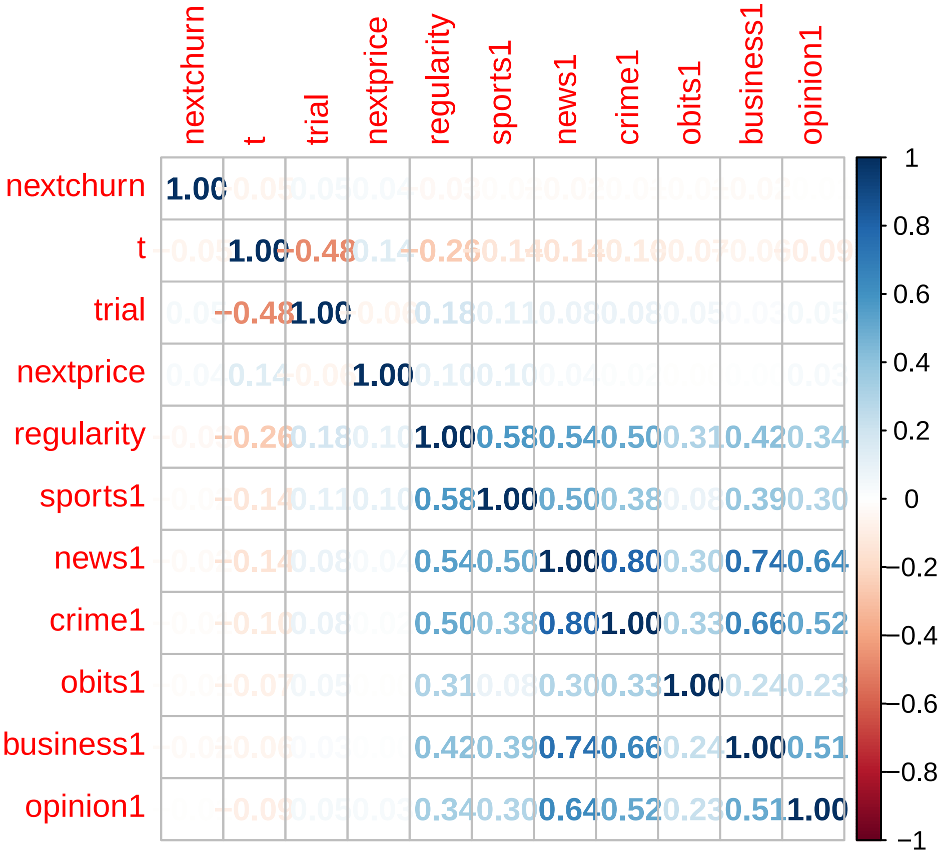
• Only sports and news content have significant impacts on churn, and both reduce the churn rate. This suggests that engaging customers with sports and news content could help retain them. The possible intention behind it is that News content is often highly relevant to users, keeping them informed about current events. Customers who find value in staying updated might engage more with the platform, leading to higher retention. Also, Sports fans often have strong loyalty to their favorite teams or leagues, driving them to consume sports content regularly.

• Other content types (crime, life, obits, business, opinion) show no statistically significant impact on churn, implying that these content types do not strongly influence customer retention. There coefficient are less interested to study.

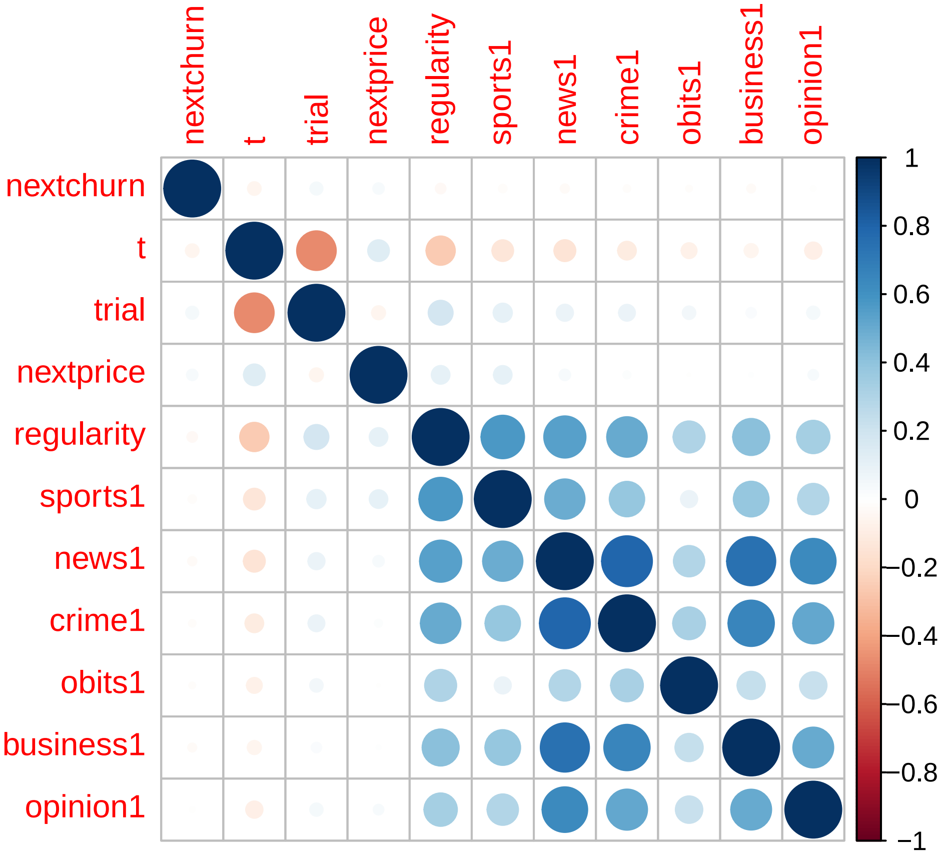
• After including regularity, we see that as the signifcance of effect that regularity brought to reduce churn rate, all content becomes not statistically significant. The effect of sports drops significantly. A possible explaination is including regularity shifts the focus from specific content categories to overall user engagement patterns, making regularity a stronger and more comprehensive predictor of churn. As a result, the impact of content variables appears diminished in the model.

• Examining vif can also give a possible explaination. After including regularity, all vif values of content rise. Especially news and lifes. From the correlation plot, we see that regularity has relatively high correlation with all content, especially sports1 and news1, which possibly explains that it reduce the apparent significance of content variables when including regularity.

```
m = cor(np[,c('nextchurn','t', "trial", "nextprice", "regularity", "sports1", "news1", "crime1", "obits
corrplot(m, method = 'number')
```


nextchurn
t
trial
nextprice
regularity
sports1
news1
crime1
obits1
business1
opinion1
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
1.00
1.00 −0.48 0.14 −0.26 0.14 0.14 0.10 0.07 0.06 0.09
−0.48 1.00 0.18 0.11 0.08 0.08 0.05 0.05
0.14 1.00 0.10 0.10 0.03
−0.26 0.18 0.10 1.00 0.58 0.54 0.50 0.31 0.42 0.34
−0.14 0.11 0.10 0.58 1.00 0.50 0.38 0.08 0.39 0.30
−0.14 0.08 0.04 0.54 0.50 1.00 0.80 0.30 0.74 0.64
−0.10 0.08 0.50 0.38 0.80 1.00 0.33 0.66 0.52
−0.07 0.05 0.31 0.08 0.30 0.33 1.00 0.24 0.23
−0.06 0.42 0.39 0.74 0.66 0.24 1.00 0.51
−0.09 0.05 0.03 0.34 0.30 0.64 0.52 0.23 0.51 1.00

```r
corrplot(m)
```



```r
lm4 <- glm(nextchurn~t+trial+nextprice+sports1+news1+crime1+life1+obits1+business1+opinion1, family = '
summary(lm4)
```

```
##
## 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
## AIC: 3162.3
##
## Number of Fisher Scoring iterations: 6
```

```
vif(lm4)
```

```
##         t     trial nextprice   sports1     news1    crime1     life1    obits1
##  1.473763  1.433924  1.034613  1.152708  3.422858  2.506937  1.848970  1.133380
## business1  opinion1
##  1.965271  1.495660
```

```
lm5 <- glm(nextchurn~t+trial+nextprice+sports1+news1+crime1+life1+obits1+business1+opinion1+regularity,
summary(lm5)
```

```
##
## 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
## AIC: 3154.4
##
## Number of Fisher Scoring iterations: 6
```

```
vif(lm5)
```

```
##          t      trial  nextprice    sports1      news1     crime1      life1
##   1.494290   1.443141   1.042231   1.758981   4.252032   2.846017   2.235592
##     obits1  business1   opinion1 regularity
##   1.196097   2.346581   1.645337   2.336147
```

4

• The results show that device usage impacts churn, with **desktop** usage significantly reducing churn (coefficient = -0.0077, p < 0.001). This suggests that users who access the platform via desktop are more likely to stay engaged, while **mobile** (-0.0021, p = 0.322) and **tablet** (-0.0065, p = 0.102) usage have no statistically significant effect on churn.

• Desktop users may engage more deeply with the platform, possibly using it for professional or extensive purposes, which contributes to retention. In contrast, mobile and tablet usage might reflect more casual engagement, explaining their weaker effects. These findings highlight the value of optimizing the desktop experience to improve retention while considering other factors like content and regularity.

• Including **t (month)**, **nextprice**, and **trial** in the model provides additional explanatory power and highlights the significance of desktop usage in reducing churn. The coefficient for **desktop** (-0.0091, p < 0.001) remains negative and becomes more significant, indicating that users who access the platform via desktop are even less likely to churn. **Tablet usage** shows a borderline significant effect (p = 0.0534), while **mobile usage** remains non-significant. Additionally, **nextprice** (0.0831, p < 0.001) and **trial** (0.3071, p = 0.0472) positively correlate with churn, confirming their roles as strong drivers of user drop-off.

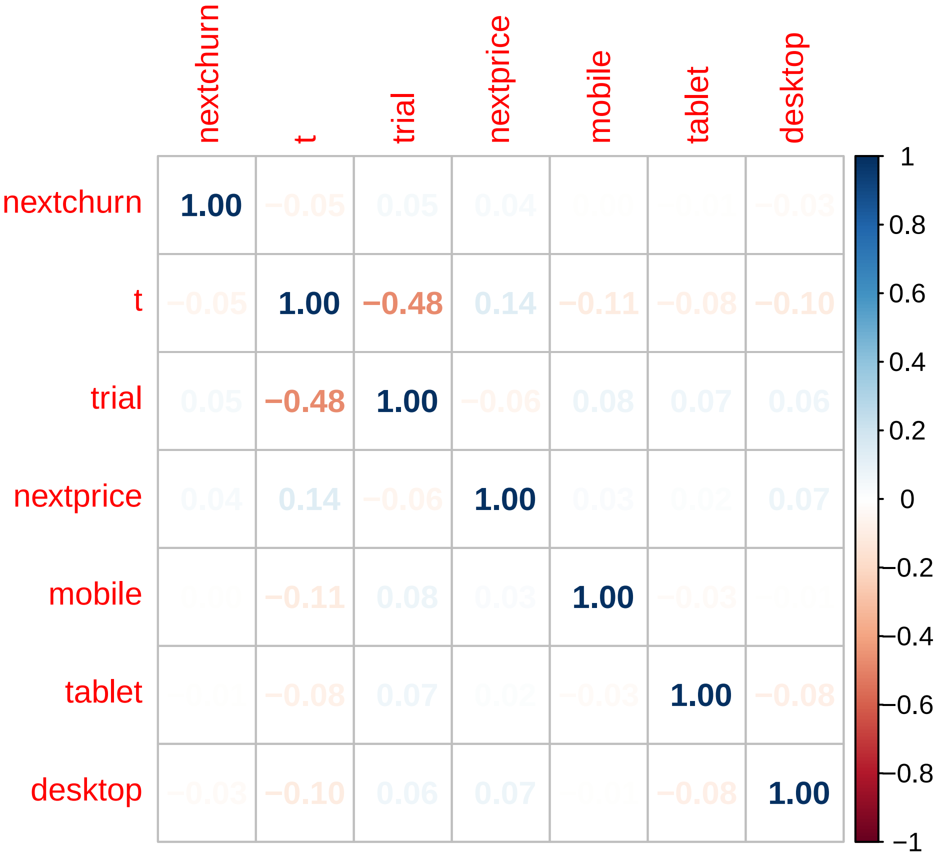
• Including these three variables likely makes desktop usage more significant because these variables account for broader temporal patterns, pricing effects, and trial-based behaviors. This reduces unexplained variability in the model, allowing the specific impact of desktop usage to emerge more clearly. Desktop users may exhibit more consistent engagement and loyalty, which stands out when other major churn predictors are accounted for. This emphasizes that desktop access is an independent factor in reducing churn.

• Observing the correlation plot the vif, we see that the correlation and vif values are really low. Specifically, the vif values of mobile, tablet, and desktop approach 0, whether or not t, mobile, and trial are included. After including these three variables, the vif values are still all less than 1.5 indicating small or ignorable multicollinearity.

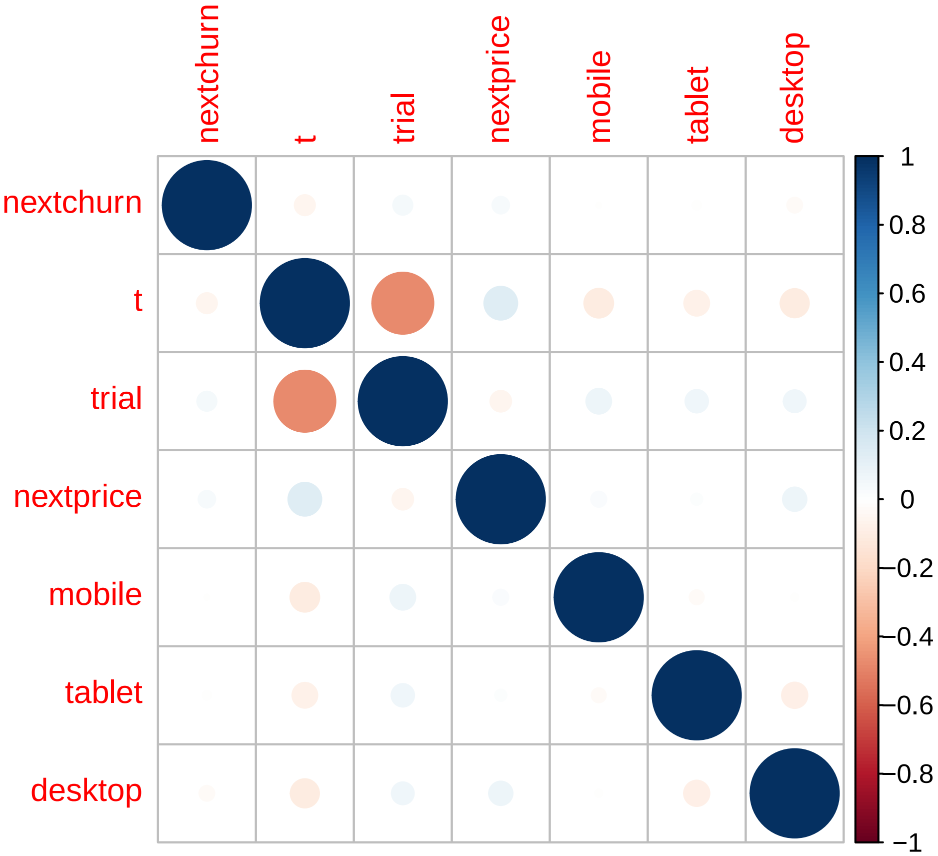
```r
head(np)
```

```
## # A tibble: 6 x 32
## # Groups:   SubscriptionId [1]
##   SubscriptionId         t churn regularity intensity sports1 news1 crime1 life1
##   <chr>              <int> <int>      <int>     <dbl>   <int> <int>  <int> <int>
## 1 0030b3e52743e6c4c~     1     0         30      39.4      27    86     35    16
## 2 0030b3e52743e6c4c~     2     0         24      19.6      21    54     16    16
## 3 0030b3e52743e6c4c~     3     0          1      18         1     4      0     0
## 4 0030b3e52743e6c4c~     4     0          3      15         0     6      1     2
## 5 0030b3e52743e6c4c~     5     0          6      17.5       0    10     11     2
## 6 0030b3e52743e6c4c~     6     0          5      12.2       2     8      5     0
## # i 23 more variables: obits1 <int>, business1 <int>, opinion1 <int>,
## #   mobile <int>, tablet <int>, desktop <int>, loc1 <int>, Loc2 <int>,
## #   Loc3 <int>, Loc4 <int>, SrcGoogle <int>, SrcDirect <int>, SrcElm <int>,
## #   SrcSocial <int>, SrcBingYahooAol <int>, SrcNewsletter <int>,
## #   SrcLegacy <int>, SrcGoogleNews <int>, SrcGoogleAd <int>, currprice <dbl>,
## #   trial <int>, nextchurn <int>, nextprice <dbl>
```

```r
m = cor(np[,c('nextchurn','t', "trial", "nextprice", "mobile", "tablet", "desktop")],  use = 'complete.
corrplot(m, method = 'number')
```


nextchurn
t
trial
nextprice
mobile
tablet
desktop
nextchurn 1.00 −0.05 0.05 0.04 0.00 −0.01 −0.03
t −0.05 1.00 −0.48 0.14 −0.11 −0.08 −0.10
trial 0.05 −0.48 1.00 −0.06 0.08 0.07 0.06
nextprice 0.04 0.14 −0.06 1.00 0.03 0.02 0.07
mobile 0.00 −0.11 0.08 0.03 1.00 −0.03 −0.01
tablet −0.01 −0.08 0.07 0.02 −0.03 1.00 −0.08
desktop −0.03 −0.10 0.06 0.07 −0.01 −0.08 1.00
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

```
corrplot(m)
```



```
lm6 <- glm(nextchurn ~ t+mobile+tablet+desktop, family = "binomial",data = np)
summary(lm6)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + mobile + tablet + desktop, family = "binomial",
##     data = np)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.516846   0.109105 -23.068  < 2e-16 ***
## t           -0.138049   0.024535  -5.627 1.84e-08 ***
## mobile      -0.002223   0.002194  -1.013 0.310920
## tablet      -0.006863   0.004048  -1.695 0.090030 .
## desktop     -0.007822   0.002219  -3.524 0.000424 ***
## ---
## 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: 3167.1  on 9530  degrees of freedom
## AIC: 3177.1
##
## Number of Fisher Scoring iterations: 6
```

```
vif(lm6)
```

```
##        t   mobile   tablet  desktop
## 1.031840 1.012326 1.014834 1.023710
```

```
lm7 <- glm(nextchurn ~ t+mobile+tablet+desktop+nextprice+trial,family = "binomial", data = np)
summary(lm7)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + mobile + tablet + desktop + nextprice +
##     trial, 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 ***
## 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 ***
## nextprice    0.083129   0.018413   4.515 6.34e-06 ***
## trial        0.307087   0.154737   1.985   0.0472 *
## ---
## 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
## AIC: 3152.5
##
## Number of Fisher Scoring iterations: 6
```

```
vif(lm7)
```

```
##         t    mobile    tablet   desktop nextprice     trial
##  1.465370  1.021615  1.019225  1.040724  1.027241  1.430038
```

```
lm9 <- glm(nextchurn ~ t+mobile+tablet+desktop+nextprice+trial+regularity+intensity,family = "binomial"
summary(lm9)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + mobile + tablet + desktop + nextprice +
##     trial + regularity + intensity, family = "binomial", data = np)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.0043050  0.3537020 -11.321  < 2e-16 ***
## t           -0.1435664  0.0292456  -4.909 9.15e-07 ***
## mobile       0.0029427  0.0028803   1.022    0.307
## tablet       0.0007470  0.0046140   0.162    0.871
## desktop     -0.0008521  0.0032701  -0.261    0.794
## nextprice    0.0879495  0.0185780   4.734 2.20e-06 ***
## trial        0.3578927  0.1563076   2.290    0.022 *
## regularity  -0.0284126  0.0110351  -2.575    0.010 *
## intensity   -0.0080202  0.0055769  -1.438    0.150
## ---
## 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: 3129.7  on 9526  degrees of freedom
## AIC: 3147.7
##
## Number of Fisher Scoring iterations: 6
```

```
vif(lm9)
```

```
##          t     mobile     tablet    desktop  nextprice      trial regularity
##   1.504734   1.869206   1.539358   2.636495   1.036134   1.457390   3.593028
##  intensity
##   1.655761
```

Content: All not important. Payment: Important. Device: Desktop important

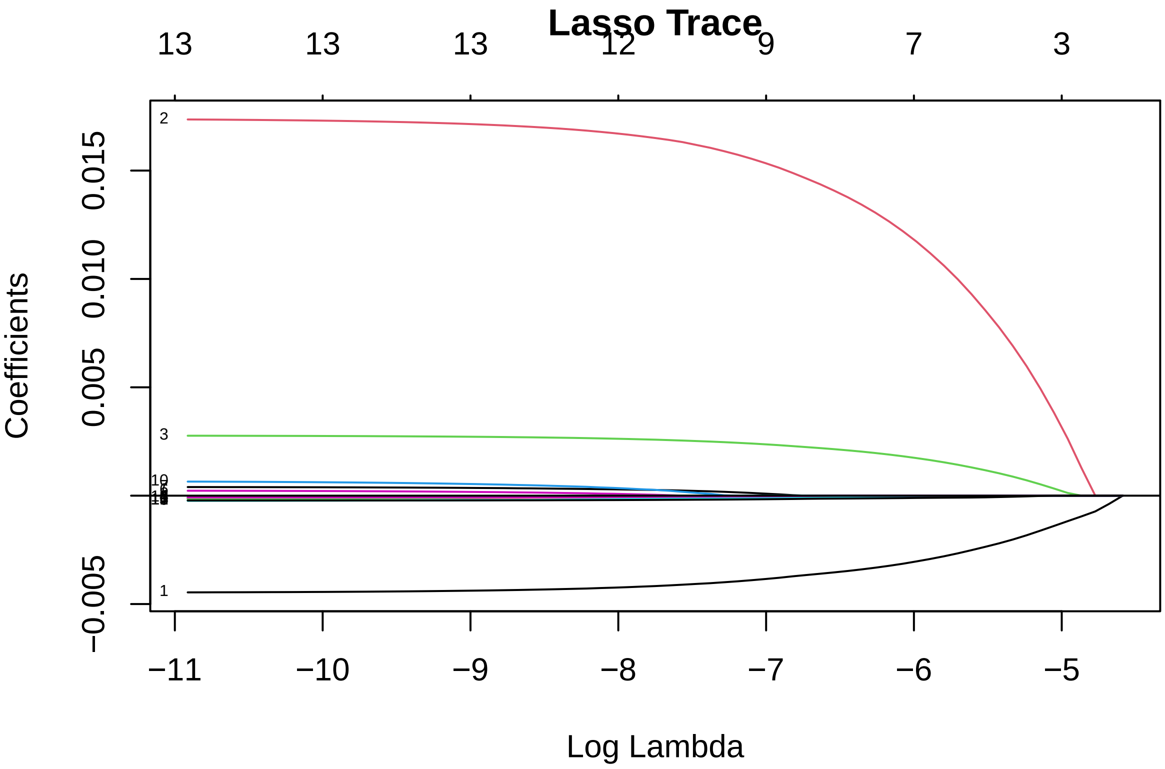
```
lm8 <- glm(nextchurn~t+trial+nextprice+sports1+news1+crime1+life1+obits1+business1+opinion1+tablet+mobil
summary(lm8)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + sports1 + news1 +
##     crime1 + life1 + obits1 + business1 + opinion1 + tablet +
##     mobile + 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
## tablet      -0.005639   0.004848  -1.163   0.2448
## mobile      -0.001506   0.003009  -0.501   0.6167
## 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
## AIC: 3161.8
##
## Number of Fisher Scoring iterations: 6
```

```
vif(lm8)
```

```
##          t     trial nextprice   sports1     news1    crime1     life1    obits1
##   1.476632  1.436740  1.036715  1.730472  4.031045  2.801223  2.023265  1.172491
## business1  opinion1    tablet    mobile   desktop
##   2.118012  1.610750  1.516447  1.821071  1.976960
```

```
x_variable = np[, c('t', 'trial', 'nextprice', "sports1", "news1", "crime1", "life1", "obits1", "busine
                    "tablet", "mobile", "desktop")]
x_variable = as.matrix(x_variable)
fit_lasso <- cv.glmnet(x_variable, np$nextchurn, alpha = 1)
plot(fit_lasso$glmnet.fit, 'lambda', label = T)
abline(h = 0)
title('Lasso Trace')
```



```
predict(fit_lasso, s = fit_lasso$lambda.min,type = 'coef')
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept)  1.618565e-02
## t           -3.720176e-03
## trial        1.488846e-02
## nextprice    2.286048e-03
## sports1     -5.242852e-05
## news1       -3.970486e-05
## crime1       .
## life1        2.307159e-05
## obits1       .
## business1    .
## opinion1     .
## tablet      -1.181741e-04
## mobile      -1.799060e-05
## desktop     -1.519113e-04
```

Relationship between engagement, device, and content (forks, pipes, and colliders)

• Device usage (e.g., desktop, tablet, mobile) and engagement metrics (regularity and intensity) behave as colliders in their relationship with churn. Both device usage and engagement directly influence churn, with engagement acting as a mediator that absorbs much of the explanatory power of device usage. When engagement metrics are included in the model, the significance of device variables diminishes or disappears, as seen with the desktop variable. This is likely because engagement metrics, which are influenced by device usage, capture the common pathway to churn, introducing spurious associations when conditioned upon. The increase in VIF values for device variables when engagement is included further supports this conclusion.

• Similarly, content variables (e.g., news, sports) and engagement also act as colliders with respect to churn. Content influences engagement, as certain types of content (e.g., sports, news) are associated with higher regularity and intensity. Both content and engagement affect churn independently, but when engagement metrics are included in the model, the effects of content variables on churn diminish significantly. This indicates that engagement mediates much of the relationship between content and churn. The rise in VIF values for content variables when engagement metrics are added suggests a correlation between these two predictors, reinforcing their role as colliders.

• In both cases, conditioning on engagement (e.g., by including regularity and intensity in the model) alters the relationships between device usage, content, and churn. These collider relationships imply that engagement metrics mediate the effects of both device usage and content on churn. As a result, spurious associations or diminished effects can arise depending on whether engagement metrics are included in the model. Understanding and properly addressing these collider dynamics is crucial for accurately interpreting the impact of device usage and content on churn.