401-final_moyi

2024-11-23

Problem 1

```
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
```

```
library(car)
```

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

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

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

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

```
library(corrplot)
```

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

```
## corrplot 0.95 loaded
```

```
# Read the data
np <- read.table("/Users/homura/Desktop/np.csv", header = TRUE, na.strings = ".", sep = " ")

# Create `nextchurn` and `nextprice` variables
np <- np %>%
  arrange(SubscriptionId, t) %>%
  group_by(SubscriptionId) %>%
  mutate(
    nextchurn = lead(churn),
    nextprice = lead(currprice),
    t = t)
```

Problem 2

Problem 2 Model 1

```
# Model 1
model1 <- glm(
  nextchurn ~ t + trial + nextprice + regularity + intensity,
  data = np,
  family = binomial
)
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
```

```
# Calculate VIFs
vif_values <- vif(model1)

# Display VIFs
print(vif_values)
```

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

Problem 2 Model 2

```
# Model 2
model2 <- glm(
  nextchurn ~ t + trial + nextprice + regularity,
  data = np,
  family = binomial
)
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
```

```
# Calculate VIFs
vif_values <- vif(model2)

# Display VIFs
print(vif_values)
```

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

Problem 2 Model 3

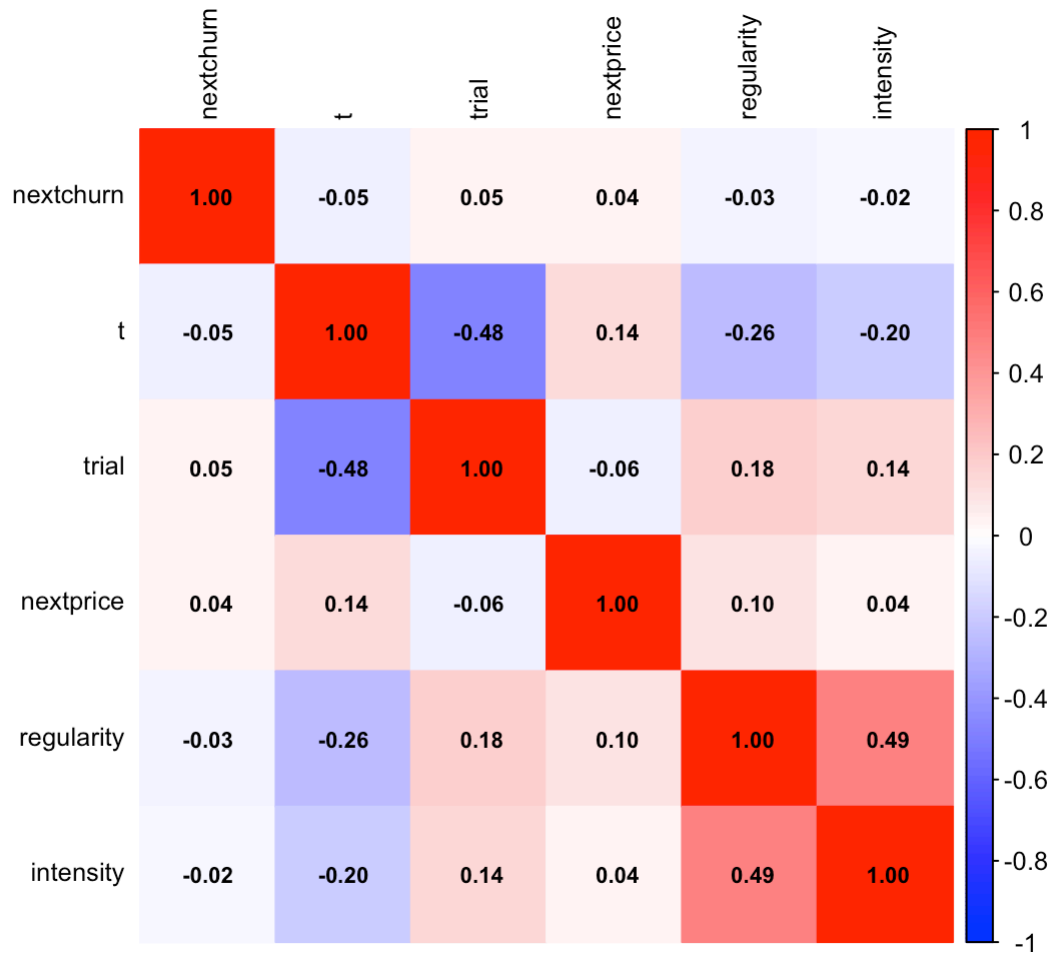
```
# Model 3
model3 <- glm(
  nextchurn ~ t + trial + nextprice + intensity,
  data = np,
  family = binomial
)
summary(model3)
```

```
##
## 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
```

```
# Calculate VIFs
vif_values <- vif(model3)

# Display VIFs
print(vif_values)
```

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

```
# Select relevant variables
selected_vars <- c("nextchurn", "t", "trial", "nextprice", "regularity", "intensity")
np_subset <- na.omit(np[, selected_vars]) # Remove rows with missing values

# Calculate the correlation matrix
cor_matrix <- cor(np_subset)

# Plot the correlation matrix
corrplot(cor_matrix, method = "color", addCoef.col = "black",
         tl.col = "black", tl.cex = 0.8, number.cex = 0.7,
         col = colorRampPalette(c("blue", "white", "red"))(200))
```



a. Looking at all three models, the trial variable shows a consistent positive and statistically significant effect on next month's churn. The coefficient for trial ranges from 0.325 to 0.360 across the models (all with $p < 0.05$). Given that most trial offers are 1 month and many customers didn't have trial offers, this positive association suggests that customers who come in through trial subscriptions are more likely to churn in their next period. This makes intuitive sense as trial subscribers may be initially attracted by the discounted rate and are more likely to cancel when faced with full-price subscriptions. This behavior pattern indicates that while trials may be effective at acquiring new customers, they might be attracting more price-sensitive subscribers who are less likely to convert to long-term customers.

b. The comparison between intensity and regularity provides interesting insights into user engagement patterns. In Model 1, which includes both variables, regularity shows a significant negative effect on churn (coefficient = -0.0265, $p < 0.001$), while intensity is not significant (coefficient = -0.00771, $p = 0.135$). When each variable is tested separately in Models 2 and 3, both become highly significant, with regularity showing a stronger effect (coefficient = -0.0319, $p < 2.08e-07$) compared to intensity (coefficient = -0.0189, $p = 0.000163$). The VIF values for both variables are relatively low (around 1.1-1.4), indicating minimal multicollinearity concerns. This suggests that regularity - the number of reading days per month - is a more reliable predictor of customer retention than intensity (page views per reading day). Organizations should therefore prioritize developing strategies that encourage consistent, regular engagement with the content rather than focusing on increasing the volume of content consumed during each visit (Model2). Regular usage habits appear to be more effective at building lasting customer relationships than intensive but potentially sporadic usage patterns.

Problem 3

```
# Load necessary libraries
library(dplyr)

# Load the dataset
np <- read.table("np.csv", header = TRUE, na.strings = ".", sep = " ")

# Group by SubscriptionId and create lead variables
np <- np %>%
  arrange(SubscriptionId, t) %>%
  group_by(SubscriptionId) %>%
  mutate(
    nextchurn = lead(churn),
    nextprice = lead(currprice)
  )

# Fit the logistic regression model for content
model_content <- glm(
  nextchurn ~ t + trial + nextprice + sports1 + news1 + crime1 + life1 + obits1 + business1 +
opinion1,
  data = np,
  family = binomial()
)

# Add regularity to the model
model_content_with_regularity <- glm(
  nextchurn ~ t + trial + nextprice + sports1 + news1 + crime1 + life1 + obits1 + business1 +
opinion1 + regularity,
  data = np,
  family = binomial()
)

# Summarize the models
summary(model_content)
```

```
##
## 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
```

```
summary(model_content_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
```

The analysis of content variables reveals that sports and news content significantly reduce churn, suggesting these categories engage users effectively and are crucial for retention strategies. However, when regularity is included in the model, these effects diminish, and regularity becomes a strong predictor of churn reduction, indicating its overarching influence on user behavior. Other content variables, such as crime, life, obits, business, and opinion, show no statistically significant impact on churn in either model. This suggests that overall user engagement, as measured by regularity, is more critical for retention than specific content categories, emphasizing the need to focus on fostering consistent user interactions.

Problem 4

```
# Fit the logistic regression model for devices
model_device <- glm(
  nextchurn ~ t + trial + nextprice + mobile + tablet + desktop,
  data = np,
  family = binomial()
)

# Summarize the model
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
```

The analysis reveals that desktop usage significantly reduces churn, with higher desktop sessions associated with improved retention, suggesting it is the most impactful device for engaging users. In contrast, mobile usage has no significant effect, and tablet usage shows a weak negative impact on churn, indicating their lesser importance. These findings highlight the need to prioritize enhancing desktop user experiences to improve retention. Additionally, other factors like trial periods and next-period pricing also influence churn, with trial users being more likely to churn and higher prices increasing churn risk. Strategies should focus on converting trial users into long-term subscribers, optimizing pricing models, and leveraging the engagement potential of desktop platforms.