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

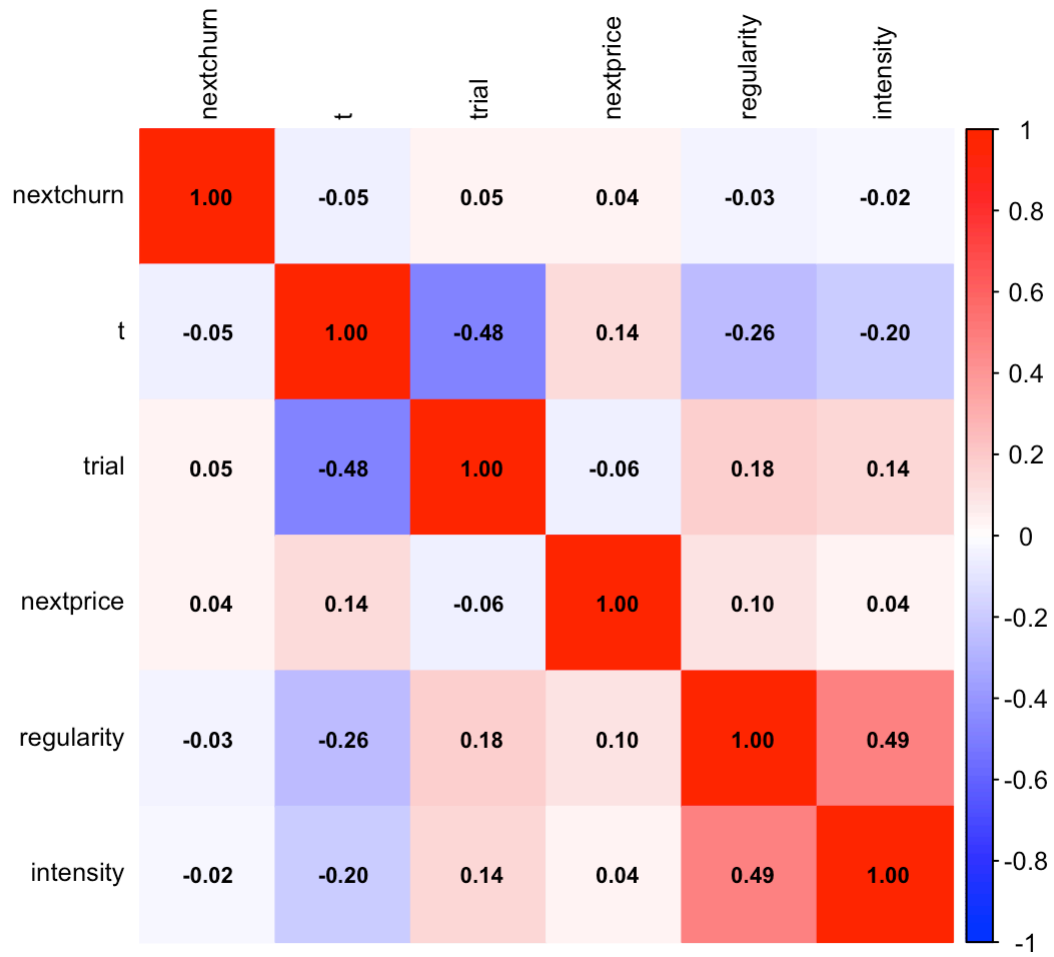
```
# Model 3
model3 <- glm(
  nextchurn ~ t + trial + nextprice + intensity,
  data = np,
  family = binomial
)
summary(model3)
```

```
##
## 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.