Assignment 4

Kristyan Dimitrov, Srividya Ganapathi, Shreyashi Ganguly, Greesham Simon, Joe Zhang

18/02/2020

Problem 1

```
library(tidyverse)
library(car)
library(glmnet)
np = read.csv('np.csv', sep = ' ',na.strings=".", stringsAsFactors=FALSE)
```

a) - Creating nextchurn & nextprice

The below code creates a column called nextchurn, which is just the churn column shifted one cell up for each customer. The empty cell left over where a new customer starts is left as an NA. In other words, in the latest month of the customer's life we don't know if he/she will churn next month. The important thing is that THIS month we have an indicator if he/she will churn next month.

```
np = np %>%
  group_by(SubscriptionId) %>%
  mutate(nextchurn = lead(churn, order_by = t))
# This creates a column called nextchurn, which is just the churn column shifted one cell up.
# The empty cell left over where a new customer starts is left as an NA i.e. in the latest month of the
```

```
np = np %>% # Same as above, but for nextprice
  group_by(SubscriptionId) %>%
  mutate(nextprice = lead(currprice, order_by = t))

np = np[c(1,2,3,(length(np)-1), 4:(length(np)-2), (length(np)))] # Reordering Columns so churn is next
```

b) Converting t to factor and displaying table(t)

```
np[['t']]= factor(np[['t']])
```

```
table(np[['t']])
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11
## 2064 1846 1663 1503 1319 1151 1009  685  489  320  121
```

c) - Running Logistic Regression

Model 1

```
logReg1 = glm(nextchurn ~ t+trial+nextprice+regularity+intensity, data = np, family = 'binomial')
summary(logReg1)
```

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

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4888  -0.3180  -0.2695  -0.2151   3.0964
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.545812   0.423913 -10.723  < 2e-16 ***
## t2           0.392508   0.283731   1.383 0.166548
## t3          -0.031308   0.296842  -0.105 0.916003
## t4           0.361422   0.292140   1.237 0.216031
## t5          -0.093176   0.310208  -0.300 0.763897
## t6          -0.270195   0.323128  -0.836 0.403050
## t7          -0.460220   0.360695  -1.276 0.201982
## t8          -0.923226   0.443729  -2.081 0.037470 *
## t9          -1.213638   0.568041  -2.137 0.032636 *
## t10         -0.292687   0.573882  -0.510 0.610043
## trial        0.803046   0.279605   2.872 0.004078 **
## nextprice    0.083950   0.018533   4.530  5.9e-06 ***
## regularity  -0.027038   0.007061  -3.829 0.000129 ***
## intensity   -0.006384   0.005112  -1.249 0.211719
## ---
## 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: 3115.9  on 9521  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3143.9
##
## Number of Fisher Scoring iterations: 6
```

```
vif(logReg1)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## t          4.934347  9        1.092730
## trial      4.665872  1        2.160063
## nextprice  1.038743  1        1.019187
## regularity 1.461983  1        1.209125
## intensity  1.439989  1        1.199996
```

Checking the Variance Inflation Factors (VIFs) None of them are larger than 10.

```
logReg2 = glm(nextchurn ~ t+trial+nextprice+regularity, data = np, family = 'binomial')
summary(logReg2)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + regularity,
##     family = "binomial", data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4773  -0.3179  -0.2709  -0.2140   3.0907
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.609117   0.421675 -10.930  < 2e-16 ***
## t2           0.427360   0.282323   1.514  0.13010
## t3           0.005646   0.295381   0.019  0.98475
## t4           0.404434   0.290162   1.394  0.16337
## t5          -0.047185   0.308139  -0.153  0.87830
## t6          -0.226755   0.321308  -0.706  0.48036
## t7          -0.419841   0.359254  -1.169  0.24255
## t8          -0.881986   0.442542  -1.993  0.04626 *
## t9          -1.171459   0.567045  -2.066  0.03884 *
## t10         -0.262583   0.573176  -0.458  0.64687
## trial        0.819507   0.279125   2.936  0.00332 **
## nextprice    0.083789   0.018510   4.527 5.99e-06 ***
## regularity  -0.031535   0.006153  -5.125 2.98e-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: 3117.7  on 9522  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3143.7
##
## Number of Fisher Scoring iterations: 6
```

Checking the Variance Inflation Factors (VIFs) None of them are larger than 10.

```
vif(logReg2)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## t          4.857849  9        1.091782
## trial      4.649945  1        2.156373
## nextprice  1.038664  1        1.019148
## regularity 1.103977  1        1.050703
```

```
logReg3 = glm(nextchurn ~ t+trial+nextprice+intensity, data = np, family = 'binomial')
summary(logReg3)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + intensity,
##     family = "binomial", data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4487  -0.3225  -0.2746  -0.2277   3.5937
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.49331    0.42109 -10.671  < 2e-16 ***
## t2           0.36733    0.28476   1.290 0.197070
## t3          -0.03467    0.29796  -0.116 0.907359
## t4           0.36581    0.29332   1.247 0.212349
## t5          -0.06471    0.31136  -0.208 0.835368
```

```
## t6          -0.21296    0.32416  -0.657 0.511204
## t7          -0.40451    0.36163  -1.119 0.263323
## t8          -0.86852    0.44438  -1.954 0.050646 .
## t9          -1.16212    0.56859  -2.044 0.040968 *
## t10         -0.21953    0.57434  -0.382 0.702291
## trial        0.74148    0.28054   2.643 0.008216 **
## nextprice    0.07537    0.01832   4.115 3.87e-05 ***
## intensity   -0.01766    0.00499  -3.538 0.000403 ***
## ---
## 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: 3131.0  on 9522  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3157
##
## Number of Fisher Scoring iterations: 6
```

Checking the Variance Inflation Factors (VIFs) None of them are larger than 10.

```
vif(logReg3)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## t         4.946304  9        1.092877
## trial     4.708602  1        2.169931
## nextprice 1.024856  1        1.012352
## intensity 1.104621  1        1.051010
```

Let's look at the correlation matrix for our predictors.

```
npSubset = np %>%  select(t, trial,nextchurn, nextprice, regularity, intensity)
```

```
## Adding missing grouping variables: `SubscriptionId`
```

```
npSubset['t'] = sapply(npSubset[['t']], as.numeric) # Converting t back to numeric for the cor() functi
cor(npSubset[c(2:length(npSubset))])
```

```
##                      t      trial nextchurn nextprice regularity  intensity
## t            1.0000000 -0.4424921        NA        NA -0.2338832 -0.1834243
## trial       -0.4424921  1.0000000        NA        NA  0.1838662  0.1494231
## nextchurn           NA         NA         1        NA         NA         NA
## nextprice           NA         NA        NA         1         NA         NA
## regularity -0.2338832  0.1838662        NA        NA  1.0000000  0.4902455
## intensity  -0.1834243  0.1494231        NA        NA  0.4902455  1.0000000
```

Let's remove the NA's from the lead variables. Note that the below has only 9,535 out of 12,170 rows from the dataset.

```
cor(na.omit(npSubset[c(2:length(npSubset))]))
```

```
##                       t       trial   nextchurn   nextprice  regularity
## t            1.00000000 -0.47865664 -0.05185924  0.13873021 -0.25766367
## trial       -0.47865664  1.00000000  0.04698046 -0.05506894  0.18447356
## nextchurn   -0.05185924  0.04698046  1.00000000  0.03573651 -0.03024197
## nextprice    0.13873021 -0.05506894  0.03573651  1.00000000  0.10124444
## regularity -0.25766367  0.18447356 -0.03024197  0.10124444  1.00000000
```

```
## intensity  -0.19693196  0.14216258 -0.02142782  0.03542222  0.48545901
##              intensity
## t          -0.19693196
## trial       0.14216258
## nextchurn  -0.02142782
## nextprice   0.03542222
## regularity  0.48545901
## intensity   1.00000000
```

There is no correlation larger than .6, which agian doesn't suggest any significant multicollinearity. It is worth pointing out, however, that there is significant collinearity b/w regularity & intensity. That's why individually, intensity and regularity appear highly significant, but when both are added to the same model only regularity is significant.

The baseline t = 1 level is the reference category for t. This means that the rest of the factors tell us how the probability of churning next month changes as time goes by FOR ALL USERS. On the other hand, trial gives us an indication how the probability that the user will churn changes after their 1-month trial ends i.e. this applies ONLY FOR USERS WHO HAD A TRIAL.

- trial - positive impact on nextchurn => customers are more likely to churn the month following their trial month
- nextprice - as price goes up customers tend to churn
- regularity and intensity - negative impact on churn

Considering the fact that trial comes up to be highly significant in predicting churn, even though few customers get the trial offer shows that customers have a very high probability of attriting after the trial period

d) Modelling with content variables

```
logRegContent1 = glm(nextchurn~t+trial+nextprice+sports1+news1+crime1+life1+obits1+business1+opinion1, 
summary(logRegContent1)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + sports1 + news1 +
##     crime1 + life1 + obits1 + business1 + opinion1, family = "binomial",
##     data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6721  -0.3221  -0.2739  -0.2230   3.7244
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.658625   0.420736 -11.073  < 2e-16 ***
## t2           0.430562   0.283814   1.517  0.12925
## t3           0.029042   0.296734   0.098  0.92203
## t4           0.431611   0.291436   1.481  0.13861
## t5          -0.009751   0.309258  -0.032  0.97485
## t6          -0.174246   0.322076  -0.541  0.58850
## t7          -0.369530   0.359960  -1.027  0.30462
## t8          -0.824530   0.443021  -1.861  0.06272 .
## t9          -1.110084   0.567423  -1.956  0.05042 .
## t10         -0.171790   0.574128  -0.299  0.76477
## trial        0.783647   0.280812   2.791  0.00526 **
## nextprice    0.079155   0.018447   4.291 1.78e-05 ***
```

```
## sports1     -0.006051   0.002526  -2.396  0.01659 *
## news1       -0.012293   0.005943  -2.069  0.03859 *
## crime1       0.008422   0.007834   1.075  0.28234
## life1        0.003183   0.008353   0.381  0.70312
## obits1      -0.009327   0.013812  -0.675  0.49951
## business1   -0.013390   0.026776  -0.500  0.61703
## opinion1     0.027934   0.027940   1.000  0.31741
## ---
## 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: 3124.0  on 9516  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3162
##
## Number of Fisher Scoring iterations: 6
```

Now we include regularity in the model

```
logRegContent2 = glm(nextchurn~t+trial+nextprice+regularity+sports1+news1+crime1+life1+obits1+business1+
summary(logRegContent2)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + regularity +
##     sports1 + news1 + crime1 + life1 + obits1 + business1 + opinion1,
##     family = "binomial", data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5294  -0.3185  -0.2709  -0.2128   3.1971
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.6155611  0.4226672 -10.920  < 2e-16 ***
## t2           0.4176536  0.2831920   1.475  0.14026
## t3          -0.0056195  0.2962624  -0.019  0.98487
## t4           0.3876065  0.2910988   1.332  0.18302
## t5          -0.0670440  0.3090762  -0.217  0.82827
## t6          -0.2474622  0.3222080  -0.768  0.44248
## t7          -0.4415044  0.3601258  -1.226  0.22021
## t8          -0.8996963  0.4432146  -2.030  0.04236 *
## t9          -1.1901965  0.5676607  -2.097  0.03602 *
## t10         -0.2734578  0.5747778  -0.476  0.63424
## trial        0.8059052  0.2801890   2.876  0.00402 **
## nextprice    0.0845657  0.0185928   4.548 5.41e-06 ***
## regularity  -0.0284847  0.0089849  -3.170  0.00152 **
## sports1     -0.0007758  0.0027521  -0.282  0.77803
## news1       -0.0083130  0.0057736  -1.440  0.14992
## crime1       0.0104412  0.0075896   1.376  0.16891
## life1        0.0041675  0.0079317   0.525  0.59929
## obits1      -0.0013451  0.0137189  -0.098  0.92190
## business1   -0.0094388  0.0265196  -0.356  0.72190
```

```
## opinion1     0.0223751  0.0270253   0.828  0.40771
## ---
## 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: 3114.3  on 9515  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3154.3
##
## Number of Fisher Scoring iterations: 6
```

Initially sports and news readers seem to have lower attrition.

After adding regularity, we see that sports1 & news1 are no longer significant. This is probably, again, due to them being correlated with regularity.

Also it shows that regularity is a stronger predictor of churn, rather than the kind of content the user reads. This seems fairly intuitive.

e) - Fitting with Location

```
logRegLocation1 = glm(nextchurn~t+trial+nextprice+loc1+Loc2+Loc3+Loc4, data = np, family = 'binomial')
summary(logRegLocation1)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + loc1 + Loc2 +
##     Loc3 + Loc4, family = "binomial", data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4182  -0.3246  -0.2730  -0.2273   3.0682
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.632407   0.418357 -11.073  < 2e-16 ***
## t2           0.435990   0.283727   1.537  0.12438
## t3           0.042774   0.296675   0.144  0.88536
## t4           0.463551   0.291143   1.592  0.11135
## t5           0.043727   0.308690   0.142  0.88735
## t6          -0.109860   0.321395  -0.342  0.73248
## t7          -0.306178   0.359308  -0.852  0.39414
## t8          -0.772394   0.442614  -1.745  0.08097 .
## t9          -1.063582   0.567137  -1.875  0.06074 .
## t10         -0.139248   0.573160  -0.243  0.80805
## trial        0.760575   0.280446   2.712  0.00669 **
## nextprice    0.073999   0.018202   4.065 4.79e-05 ***
## loc1        -0.004907   0.002014  -2.436  0.01485 *
## Loc2        -0.007137   0.005967  -1.196  0.23167
## Loc3        -0.005230   0.005573  -0.938  0.34804
## Loc4        -0.006250   0.005082  -1.230  0.21882
## ---
## 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 9519  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3166
##
## Number of Fisher Scoring iterations: 6
```

```
logRegLocation2 = glm(nextchurn~t+trial+nextprice+regularity+loc1+Loc2+Loc3+Loc4, data = np, family = '
summary(logRegLocation2)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + regularity +
##     loc1 + Loc2 + Loc3 + Loc4, family = "binomial", data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4772  -0.3179  -0.2711  -0.2138   3.0776
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.610e+00  4.217e-01 -10.930  < 2e-16 ***
## t2           4.275e-01  2.824e-01   1.514  0.13004
## t3           5.670e-03  2.954e-01   0.019  0.98469
## t4           4.044e-01  2.902e-01   1.394  0.16343
## t5          -4.703e-02  3.082e-01  -0.153  0.87870
## t6          -2.265e-01  3.213e-01  -0.705  0.48099
## t7          -4.199e-01  3.593e-01  -1.169  0.24246
## t8          -8.824e-01  4.426e-01  -1.994  0.04618 *
## t9          -1.172e+00  5.671e-01  -2.066  0.03880 *
## t10         -2.640e-01  5.733e-01  -0.460  0.64518
## trial        8.193e-01  2.793e-01   2.933  0.00335 **
## nextprice    8.380e-02  1.852e-02   4.526 6.00e-06 ***
## regularity  -3.126e-02  8.000e-03  -3.907 9.35e-05 ***
## loc1         3.926e-05  2.170e-03   0.018  0.98557
## Loc2        -1.247e-03  5.926e-03  -0.210  0.83335
## Loc3         3.862e-04  5.407e-03   0.071  0.94305
## Loc4        -3.223e-04  5.058e-03  -0.064  0.94919
## ---
## 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: 3117.6  on 9518  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3151.6
##
## Number of Fisher Scoring iterations: 6
```

Previously it showed that loc1 seems associated with lower attrition. However, after adding regularity, we observe that loc1 is no longer significant. This probably makes sense: most users have a specific location

where they read 'most regularly'. As a result, the two variables would be highly correlated. Again we see that when used alone, location is also a proxy for regularity, which is why including regularity itself reduces importance of location as also improves predictive power of the model

f) - Effects of Source on Churn

```
logRegSource1 = glm(nextchurn~t+trial+nextprice+SrcGoogle+SrcDirect+SrcElm+SrcSocial+SrcBingYahooAol+Sr
summary(logRegSource1)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + SrcGoogle +
##     SrcDirect + SrcElm + SrcSocial + SrcBingYahooAol + SrcNewsletter +
##     SrcLegacy + SrcGoogleNews + SrcGoogleAd, family = "binomial",
##     data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4931  -0.3263  -0.2734  -0.2243   3.0446
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.573124   0.420980 -10.863  < 2e-16 ***
## t2               0.440265   0.282905   1.556  0.11965
## t3               0.055030   0.295791   0.186  0.85241
## t4               0.481331   0.290381   1.658  0.09740 .
## t5               0.058700   0.308063   0.191  0.84888
## t6              -0.096994   0.320823  -0.302  0.76240
## t7              -0.292995   0.358722  -0.817  0.41406
## t8              -0.754859   0.442183  -1.707  0.08780 .
## t9              -1.051003   0.567255  -1.853  0.06391 .
## t10             -0.116211   0.572977  -0.203  0.83928
## trial            0.762053   0.279598   2.726  0.00642 **
## nextprice        0.072072   0.018348   3.928 8.56e-05 ***
## SrcGoogle       -0.004552   0.002459  -1.851  0.06411 .
## SrcDirect       -0.009186   0.006367  -1.443  0.14908
## SrcElm          -0.015924   0.019650  -0.810  0.41773
## SrcSocial       -0.006519   0.005505  -1.184  0.23634
## SrcBingYahooAol -0.009433   0.007216  -1.307  0.19113
## SrcNewsletter   -0.016405   0.009127  -1.798  0.07225 .
## SrcLegacy       -0.002483   0.007513  -0.331  0.74097
## SrcGoogleNews   -0.050155   0.051334  -0.977  0.32856
## SrcGoogleAd      0.005598   0.009569   0.585  0.55857
## ---
## 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: 3127.7  on 9514  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3169.7
##
## Number of Fisher Scoring iterations: 7
```

Now we add the regularity variable

```
logRegSource2 = glm(nextchurn~t+trial+nextprice+regularity+SrcGoogle+SrcDirect+SrcElm+SrcSocial+SrcBing
summary(logRegSource2)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + regularity +
##     SrcGoogle + SrcDirect + SrcElm + SrcSocial + SrcBingYahooAol +
##     SrcNewsletter + SrcLegacy + SrcGoogleNews + SrcGoogleAd,
##     family = "binomial", data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6655  -0.3183  -0.2710  -0.2144   3.0898
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.5679638  0.4228406 -10.803  < 2e-16 ***
## t2               0.4256822  0.2824191   1.507 0.131741
## t3               0.0076166  0.2955075   0.026 0.979437
## t4               0.4084089  0.2904834   1.406 0.159735
## t5              -0.0424887  0.3085703  -0.138 0.890481
## t6              -0.2201784  0.3218100  -0.684 0.493857
## t7              -0.4141057  0.3596262  -1.151 0.249531
## t8              -0.8751400  0.4429479  -1.976 0.048187 *
## t9              -1.1670420  0.5678074  -2.055 0.039845 *
## t10             -0.2544152  0.5737429  -0.443 0.657454
## trial            0.8142712  0.2793537   2.915 0.003559 **
## nextprice        0.0818112  0.0186196   4.394 1.11e-05 ***
## regularity      -0.0314403  0.0087537  -3.592 0.000329 ***
## SrcGoogle        0.0011757  0.0027740   0.424 0.671702
## SrcDirect       -0.0009200  0.0063625  -0.145 0.885023
## SrcElm          -0.0077561  0.0195982  -0.396 0.692286
## SrcSocial       -0.0003421  0.0054437  -0.063 0.949898
## SrcBingYahooAol -0.0023695  0.0068310  -0.347 0.728681
## SrcNewsletter   -0.0068470  0.0090915  -0.753 0.451379
## SrcLegacy        0.0025654  0.0072081   0.356 0.721907
## SrcGoogleNews   -0.0386299  0.0496324  -0.778 0.436380
## SrcGoogleAd      0.0115647  0.0091796   1.260 0.207732
## ---
## 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: 3113.9  on 9513  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3157.9
##
## Number of Fisher Scoring iterations: 7
```

None of the sources have a particularly significant impact on attrition. Again, the different source variables become less significant after we add regularity. This also makes sense - they are correlated with regularity, probably because people tend to visit a given website via the same source.

g) Effects of Device on Churn

```
logRegDevice1 = glm(nextchurn~t+trial+nextprice+mobile+tablet+desktop, data = np, family = 'binomial')
summary(logRegDevice1)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + mobile + tablet +
##     desktop, family = "binomial", data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4447  -0.3190  -0.2745  -0.2192   3.1705
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.622714   0.420265 -11.000  < 2e-16 ***
## t2           0.422335   0.283244   1.491   0.1359
## t3           0.020979   0.296119   0.071   0.9435
## t4           0.428524   0.290813   1.474   0.1406
## t5          -0.007872   0.308554  -0.026   0.9796
## t6          -0.173408   0.321401  -0.540   0.5895
## t7          -0.365643   0.359302  -1.018   0.3088
## t8          -0.827883   0.442596  -1.871   0.0614 .
## t9          -1.116930   0.567123  -1.969   0.0489 *
## t10         -0.174344   0.573330  -0.304   0.7611
## trial        0.774195   0.280042   2.765   0.0057 **
## nextprice    0.079265   0.018392   4.310 1.63e-05 ***
## mobile      -0.003109   0.002245  -1.385   0.1661
## tablet      -0.007695   0.004051  -1.900   0.0575 .
## desktop     -0.008930   0.002283  -3.912 9.15e-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: 3122.2  on 9520  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3152.2
##
## Number of Fisher Scoring iterations: 6
```

As before, we add regularity.

```
logRegDevice2 = glm(nextchurn~t+trial+nextprice+regularity+mobile+tablet+desktop, data = np, family = '
summary(logRegDevice2)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + regularity +
##     mobile + tablet + desktop, family = "binomial", data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -0.4756  -0.3181  -0.2715  -0.2135   3.0981
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.6066114  0.4215649 -10.927  < 2e-16 ***
## t2           0.4188812  0.2826984   1.482  0.13841
## t3          -0.0020631  0.2957258  -0.007  0.99443
## t4           0.3961066  0.2905421   1.363  0.17278
## t5          -0.0539372  0.3084945  -0.175  0.86121
## t6          -0.2334575  0.3216630  -0.726  0.46797
## t7          -0.4243097  0.3595515  -1.180  0.23796
## t8          -0.8864489  0.4428030  -2.002  0.04530 *
## t9          -1.1743362  0.5672540  -2.070  0.03843 *
## t10         -0.2516388  0.5737019  -0.439  0.66093
## trial        0.8056843  0.2796936   2.881  0.00397 **
## nextprice    0.0839664  0.0185353   4.530  5.9e-06 ***
## regularity  -0.0281926  0.0109673  -2.571  0.01015 *
## mobile       0.0016174  0.0027681   0.584  0.55900
## tablet      -0.0009551  0.0045151  -0.212  0.83248
## desktop     -0.0024752  0.0030563  -0.810  0.41800
## ---
## 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: 3115.7  on 9519  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3147.7
##
## Number of Fisher Scoring iterations: 6
```

Desktop users have a significantly lower probability of attrition.

Similarly to before, we see that desktop is no longer significant, upon adding regularity.

Overall, we can conclude that device (except Desktop), location, source, and content don't have a significant impact on churn next month.

h) Fitting with all variables at the same time

Without regularity

```
logRegAllVar1 = glm(nextchurn~t+trial+nextprice+sports1+news1+crime1+life1+obits1+business1+opinion1+mo
summary(logRegAllVar1)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + sports1 + news1 +
##     crime1 + life1 + obits1 + business1 + opinion1 + mobile +
##     tablet + desktop + loc1 + Loc2 + Loc3 + Loc4 + SrcGoogle +
##     SrcDirect + SrcElm + SrcSocial + SrcBingYahooAol + SrcNewsletter +
##     SrcLegacy + SrcGoogleNews + SrcGoogleAd, family = "binomial",
##     data = np)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -0.6132  -0.3177  -0.2752  -0.2113   3.6240
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.595736   0.423161 -10.860  < 2e-16 ***
## t2               0.420422   0.284020   1.480  0.13881
## t3               0.013293   0.296959   0.045  0.96430
## t4               0.416420   0.291915   1.427  0.15372
## t5              -0.029260   0.309771  -0.094  0.92475
## t6              -0.196076   0.322626  -0.608  0.54335
## t7              -0.392818   0.360400  -1.090  0.27573
## t8              -0.849940   0.443539  -1.916  0.05533 .
## t9              -1.140817   0.568332  -2.007  0.04472 *
## t10             -0.191957   0.574706  -0.334  0.73837
## trial            0.773725   0.281074   2.753  0.00591 **
## nextprice        0.079274   0.018619   4.258 2.07e-05 ***
## sports1         -0.001871   0.003185  -0.587  0.55704
## news1           -0.010945   0.006341  -1.726  0.08431 .
## crime1           0.010670   0.007996   1.334  0.18206
## life1            0.004724   0.008519   0.555  0.57920
## obits1          -0.003434   0.014558  -0.236  0.81355
## business1       -0.003684   0.026973  -0.137  0.89137
## opinion1         0.029384   0.028368   1.036  0.30028
## mobile          -0.004627   0.005197  -0.890  0.37336
## tablet          -0.008480   0.006389  -1.327  0.18440
## desktop         -0.010031   0.005265  -1.905  0.05675 .
## loc1             0.002952   0.003871   0.762  0.44580
## Loc2            -0.002579   0.006908  -0.373  0.70892
## Loc3             0.002423   0.006346   0.382  0.70259
## Loc4             0.002052   0.005740   0.357  0.72079
## SrcGoogle        0.002795   0.004355   0.642  0.52101
## SrcDirect        0.000122   0.007291   0.017  0.98665
## SrcElm          -0.011243   0.020487  -0.549  0.58315
## SrcSocial        0.002140   0.006597   0.324  0.74558
## SrcBingYahooAol -0.001601   0.007889  -0.203  0.83916
## SrcNewsletter   -0.007362   0.009845  -0.748  0.45459
## SrcLegacy        0.002764   0.009129   0.303  0.76205
## SrcGoogleNews   -0.036289   0.050144  -0.724  0.46925
## SrcGoogleAd      0.015687   0.010143   1.547  0.12195
## ---
## 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: 3111.8  on 9500  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3181.8
##
## Number of Fisher Scoring iterations: 7
```

With regularity

```r
logRegAllVar2 = glm(nextchurn~t+trial+nextprice+regularity+sports1+news1+crime1+life1+obits1+business1+
summary(logRegAllVar2)
```

```
##
## Call:
## glm(formula = nextchurn ~ t + trial + nextprice + regularity +
##     sports1 + news1 + crime1 + life1 + obits1 + business1 + opinion1 +
##     mobile + tablet + desktop + loc1 + Loc2 + Loc3 + Loc4 + SrcGoogle +
##     SrcDirect + SrcElm + SrcSocial + SrcBingYahooAol + SrcNewsletter +
##     SrcLegacy + SrcGoogleNews + SrcGoogleAd, family = "binomial",
##     data = np)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6232  -0.3196  -0.2707  -0.2114   3.3750
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.5785854  0.4237442 -10.805  < 2e-16 ***
## t2               0.4142782  0.2836187   1.461  0.14410
## t3              -0.0069695  0.2967389  -0.023  0.98126
## t4               0.3886750  0.2918150   1.332  0.18289
## t5              -0.0656761  0.3098351  -0.212  0.83213
## t6              -0.2427331  0.3229920  -0.752  0.45234
## t7              -0.4387330  0.3607318  -1.216  0.22390
## t8              -0.8963209  0.4438419  -2.019  0.04344 *
## t9              -1.1877886  0.5685850  -2.089  0.03671 *
## t10             -0.2608919  0.5756243  -0.453  0.65038
## trial            0.7932627  0.2807904   2.825  0.00473 **
## nextprice        0.0827625  0.0187097   4.424 9.71e-06 ***
## regularity      -0.0239147  0.0118439  -2.019  0.04347 *
## sports1         -0.0002774  0.0031158  -0.089  0.92907
## news1           -0.0104378  0.0062063  -1.682  0.09260 .
## crime1           0.0111972  0.0079042   1.417  0.15660
## life1            0.0037999  0.0083598   0.455  0.64943
## obits1           0.0004074  0.0144241   0.028  0.97747
## business1       -0.0035886  0.0267234  -0.134  0.89318
## opinion1         0.0273668  0.0280416   0.976  0.32910
## mobile          -0.0007947  0.0053501  -0.149  0.88191
## tablet          -0.0036682  0.0065805  -0.557  0.57723
## desktop         -0.0047987  0.0056598  -0.848  0.39651
## loc1             0.0021552  0.0037426   0.576  0.56471
## Loc2            -0.0023845  0.0066659  -0.358  0.72056
## Loc3             0.0014169  0.0061245   0.231  0.81705
## Loc4             0.0013998  0.0055536   0.252  0.80100
## SrcGoogle        0.0024434  0.0042094   0.580  0.56161
## SrcDirect        0.0009272  0.0070810   0.131  0.89582
## SrcElm          -0.0095933  0.0202406  -0.474  0.63553
## SrcSocial        0.0018560  0.0063978   0.290  0.77175
## SrcBingYahooAol -0.0015209  0.0074890  -0.203  0.83907
## SrcNewsletter   -0.0059039  0.0095992  -0.615  0.53853
## SrcLegacy        0.0022833  0.0088296   0.259  0.79595
## SrcGoogleNews   -0.0353034  0.0494388  -0.714  0.47518
## SrcGoogleAd      0.0146511  0.0099414   1.474  0.14055
```

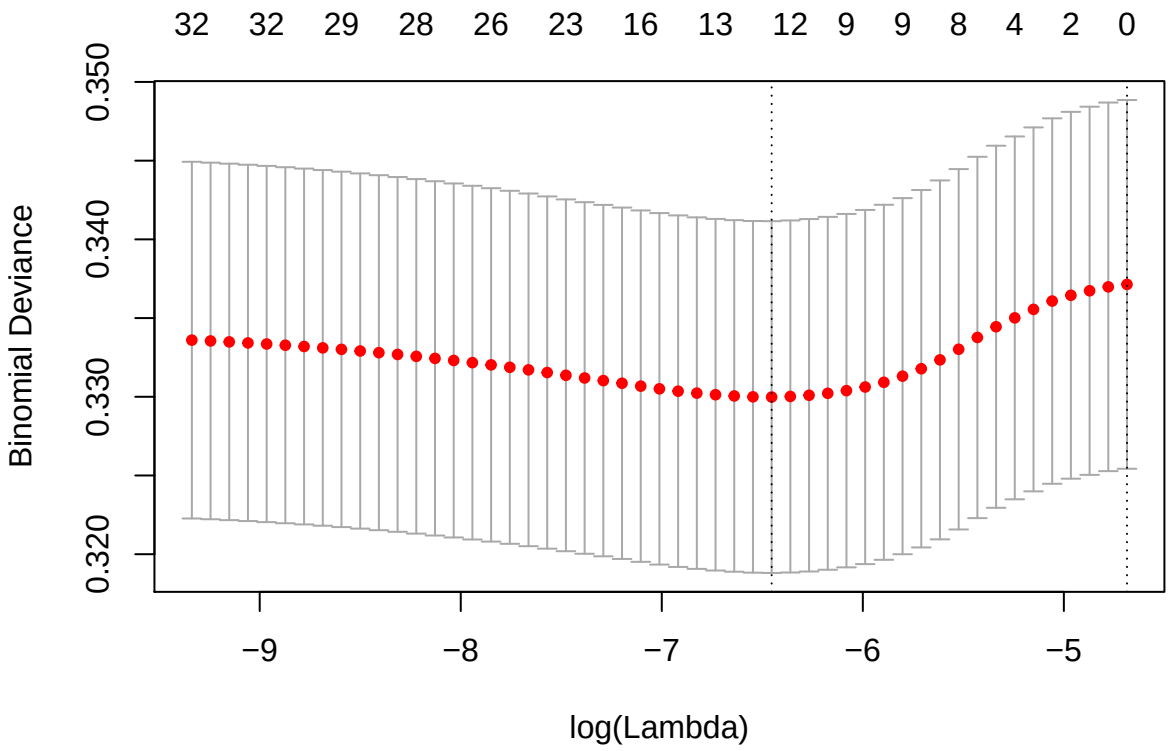
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3212.9  on 9534  degrees of freedom
## Residual deviance: 3107.8  on 9499  degrees of freedom
##   (2635 observations deleted due to missingness)
## AIC: 3179.8
##
## Number of Fisher Scoring iterations: 7
```

Most of the variables lose significance when included all at once in the model except for trial, nextprice and regularity which continue to be significant.

Let's try to use Lasso with Cross-Validation.

```
# Setting up data
set.seed(123456789)
y = na.omit(npSubset)$nextchurn
x = cbind(np[, c(2, 4)], scale(np[, c(5,7:29,31:32)])) # Scaling (aka standardizing) only the numeric v
xx = model.matrix(nextchurn ~ ., data = x)
lassoCV = cv.glmnet(xx,y,alpha=1,nfold=10, family = 'binomial')
```

```
plot(lassoCV)
```



This is our optimal lambda from the plot above.

```
lassoCV$lambda.min
```

```
## [1] 0.001574988
```

These are the coefficients for the Logistic Regression using the best lambda.

```
coef(lassoCV$glmnet.fit)[, which(lassoCV$lambda == lassoCV$lambda.min)]
```

```
##     (Intercept)     (Intercept)              t2              t3
##    -3.27287029      0.00000000      0.33100117      0.00000000
##              t4              t5              t6              t7
##     0.33973806      0.00000000     -0.03455311     -0.16832297
##              t8              t9             t10             t11
##    -0.48335903     -0.59823105      0.00000000      0.00000000
##      regularity         sports1          news1          crime1
##    -0.17852609      0.00000000      0.00000000      0.00000000
##           life1          obits1       business1        opinion1
##     0.00000000      0.00000000      0.00000000      0.00000000
##          mobile          tablet         desktop            loc1
##     0.00000000      0.00000000     -0.05254505      0.00000000
##            Loc2            Loc3            Loc4       SrcGoogle
##     0.00000000      0.00000000      0.00000000      0.00000000
##       SrcDirect          SrcElm       SrcSocial SrcBingYahooAol
##     0.00000000      0.00000000      0.00000000      0.00000000
##   SrcNewsletter       SrcLegacy   SrcGoogleNews     SrcGoogleAd
##    -0.01309199      0.00000000      0.00000000      0.01806412
##          trial       nextprice
##     0.22382809      0.21031317
```

We notice that all of the variables that we previously showed as insignificant (device, content, source, location) have mostly been shrunk to 0. In other words, we once again show that these variables don't bring any significant information. Significant variables are Desktop, regularity, SrcNewsletter, SrcGoogleAd, trial and nextprice, besides some levels of t.

i)

- No association with churn
  - We are not able to conclude that the source, location, device, or content that the user consumes are predictive or associated with their probability of churn, except for Desktop (negative impact on churn).
- Strong drivers of churn (do less of these things)
  - The end of the trial and the next prince that the user will have to pay appears to be a driver of churn. Some Customer Life-Time Value analysis would be appropriate to find out if reducing the subscription cost or making the the trial longer would increase Customer Equity.
- Strong drivers of retention (do more of these things)
  - Regularity appears to be a meaningful driver of retention. Engaging the users, for example via personalized emails with content appropriate for them might be a good idea to boost regularity.
- Questionable drivers of churn
  - t8 & t9 - appear to be mildly significant. It might be a good idea to check if there is an increase in the harard rate at that time through some Survival Analysis. It might be that users just get bored with our content. Again, engagement and personalization might be appropriate
- Questionable drivers of retention
  - In some of our regressios, the news content variable appeared slightly significant. We suspect that news content is the main driver in user regularity (i.e. people read the news every day). Therefore, making sure that our news experience is seamless and personalized might be a driver of retention, but something, which would need more analysis and investigation. Also some sources appear significant in Lasso, this also might require further testing.

Problem 2

$$n_1' = n_0'P + a'$$
$$n_2' = n_1'P + a' = (n_0'P + a') \times P + a' = n_0'P^2 + a'(P + I)$$
$$n_3' = n_2'P + a' = n_0'P^3 + a'(P^2 + P + I)$$
$$n_t' = n_0'P^t + a'(P^{t-1} + P^{t-2} + ...P + I)$$
$$\text{Let } m = P^{t-1} + P^{t-2} + ...P + I \quad (1)$$
$$Pm = P^t + P^{t-1} + ... + P^2 + P \quad (2)$$
$$(2) - (1) : (I - P) \times m = I - P^t$$
$$m = \frac{I - P^t}{I - P} = (I - P^t)(I - P)^{-1}$$
$$\text{Thus}, n_t' = n_0'P^t + a'm = n_0'P^t + a'(I - P^t)(I - P)^{-1}$$

Problem 3

Part (a): Transition Probability Matrix

```
news <- read.csv("/Users/shreyashiganguly/Documents/Northwestern_MSiA/Winter 2020/Data Mining/CLV & CRM/
colnames(news) <- c("ProsL","ProsH","TrialL","TrialH","FullL","FullH","ChurnL","ChurnH","PageView")

#Transition Probability Matrix
P <- matrix(0, nrow=8, ncol=8)
P <- news[-c(9)]/rowSums(news[-c(9)])
print("The transition probability matrix is as follows:")
```

```
## [1] "The transition probability matrix is as follows:"
```

```
print(P)
```

```
##            ProsL      ProsH       TrialL      TrialH       FullL
## ProsL  0.6360759 0.08016878 0.0875527426 0.092827004 0.052742616
## ProsH  0.1660978 0.60750853 0.0170648464 0.150170648 0.007963595
## TrialL 0.0000000 0.00000000 0.0816696915 0.030852995 0.560798548
## TrialH 0.0000000 0.00000000 0.0100896861 0.015695067 0.150224215
## FullL  0.0000000 0.00000000 0.0003869969 0.000000000 0.833978328
## FullH  0.0000000 0.00000000 0.0000000000 0.001325256 0.180802726
## ChurnL 0.0000000 0.00000000 0.0063784550 0.002126152 0.012756910
## ChurnH 0.0000000 0.00000000 0.0067226891 0.010084034 0.001680672
##              FullH      ChurnL      ChurnH
## ProsL  0.041139241 0.008438819 0.001054852
## ProsH  0.046643914 0.001137656 0.003412969
## TrialL 0.266787659 0.047186933 0.012704174
## TrialH 0.774663677 0.013452915 0.035874439
## FullL  0.118808050 0.043150155 0.003676471
## FullH  0.785687240 0.009844756 0.022340023
## ChurnL 0.002834869 0.918497519 0.057406095
## ChurnH 0.010084034 0.258823529 0.712605042
```

Part (b): Value Vector

```
sub = c(0,0,1,1,10,10,0,0)
page = news$PageView
v <- matrix(0, nrow=8, ncol=1)
for (i in 1:8) {
    v[i,1] <- sub[i] + 0.002*page[i]
}
print("The value vector is as follows:")
```

```
## [1] "The value vector is as follows:"
```

```
print(v)
```

```
##         [,1]
## [1,]  0.0154
## [2,]  0.7606
## [3,]  1.0268
## [4,]  1.5570
## [5,] 10.0084
## [6,] 10.5014
## [7,]  0.0042
## [8,]  0.3272
```

Part (c): Customer Equity (lifetime)

```
n <- c(5000,5000,1000,1000,3000,3000,4000,4000)
CE <- t(n) %*% solve(diag(rep(1,8)) - P/1.01) %*% v
sprintf("Customer Equity = $%f",CE)
```

```
## [1] "Customer Equity = $10902522.365336"
```

Part (d): Customer Equity upon cutting advertisements in half

```
v_new <- matrix(0, nrow=8, ncol=1)
for (i in 1:8) {
    v_new[i,1] <- sub[i] + 0.001*page[i]
}

CE_new <- t(n) %*% solve(diag(rep(1,8)) - P/1.01) %*% v_new
sprintf("Customer Equity on cutting advertisements in half = $%f",CE_new)
```

```
## [1] "Customer Equity on cutting advertisements in half = $10730530.535510"
```

```
sprintf("CE reduces by = $%f",abs(CE_new - CE))
```

```
## [1] "CE reduces by = $171991.829826"
```

Part (e): Projecting cash flows for next 36 months

```
CE_36 <- 0
n_t <- n
for (t in 1:36) {
    n_t <- n_t %*% as.matrix(P)
    CE_t <- n_t %*% v_new/(1.01)**t
    CE_36 <- CE_36 + CE_t
```

```
}
sprintf("Customer Equity till 36 months = $%f",CE_36)
```

```
## [1] "Customer Equity till 36 months = $3745127.035670"
```

Part (f): Change in CE upon acquiring new prospects

```
a <- c(100,100,0,0,0,0,0,0)
CE_36_a <- 0
n_t <- n
for (t in 1:36) {
    n_t <- n_t %*% as.matrix(P) + a
    CE_t <- n_t %*% v_new/(1.01)**t
    CE_36_a <- CE_36_a + CE_t
}
sprintf("Customer Equity till 36 months = $%f",CE_36_a)
```

```
## [1] "Customer Equity till 36 months = $4317320.886825"
```

```
sprintf("Expected number of customers (trial + full) = %f",sum(n_t[c(3,4,5,6)]))
```

```
## [1] "Expected number of customers (trial + full) = 14708.508158"
```

```
sprintf("Increase in CE = $%f", CE_36_a - CE_36)
```

```
## [1] "Increase in CE = $572193.851155"
```

Part (g): Increased retention upon reducing advertisements

```
P_new <- P
P_new[1:2,1:2] <- P[1:2,1:2]-0.01
P_new[3:6,7:8] <- P[3:6,7:8]-0.01
P_new[1:2,3:4] <- P[1:2,3:4]+0.01
P_new[3:6,5:6] <- P[3:6,5:6]+0.01
CE_36_r <- 0
n_t <- n
for (t in 1:36) {
    n_t <- n_t %*% as.matrix(P_new) + a
    CE_t <- n_t %*% v_new/(1.01)**t
    CE_36_r <- CE_36_r + CE_t
}
sprintf("Customer Equity till 36 months = %f",CE_36_r)
```

```
## [1] "Customer Equity till 36 months = 5303799.969937"
```

```
sprintf("Expected number of customers (trial + full) = %f",sum(n_t[c(3,4,5,6)]))
```

```
## [1] "Expected number of customers (trial + full) = 19949.271589"
```

```
sprintf("Increase in CE = %f", CE_36_r - CE_36_a)
```

```
## [1] "Increase in CE = 986479.083112"
```