Question 3

MGSC 410, Fall 2020, Professor Houldsworth

Timothy Tan

```
limited_version <- read.csv("binary_limited.csv")
view(limited_version)
str(limited_version)
## 'data.frame':    32066 obs. of  28 variables:
##  $ ID                     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Language               : chr  "POR" "EBR" "ESP" "KOR" ...
##  $ Subscription.Type      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Subscription.Event.Type: int  0 0 0 0 0 0 1 0 0 1 ...
##  $ Purchase.Store         : int  0 1 1 0 0 1 1 1 1 0 ...
##  $ Puchase.Amounts        : chr  "#VALUE!" "39" "0" "#VALUE!" ...
##  $ Subscription.Start.Date: chr  "12/28/2018" "11/28/2019" "12/31/2018"
"11/7/2019" ...
##  $ Subscription.Expiration: chr  "6/28/2019" "2/28/2020" "12/31/2019"
"2/7/2020" ...
##  $ Duration               : int  182 92 365 92 92 92 113 92 97 350 ...
##  $ Demo.User              : int  1 0 0 1 0 1 1 1 0 0 ...
##  $ Free.Trial.User        : int  0 0 0 0 0 0 0 0 1 1 ...
##  $ Auto.Renew             : int  0 0 0 0 0 0 0 1 1 1 ...
##  $ Country                : chr  "US/Canada" "Other" "US/Canada"
"US/Canada" ...
##  $ User.Type              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Lead.Platform..App...0.: int  0 1 1 0 1 0 0 0 1 0 ...
##  $ Email.Subscriber       : int  1 0 1 1 1 1 1 1 0 0 ...
##  $ Push.Notifications     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Send.Count             : int  63 4 1 14 80 162 2 25 52 1 ...
##  $ Open.Count             : int  7 3 0 0 5 1 0 17 11 0 ...
##  $ Click.Count            : int  0 0 0 0 1 0 0 4 0 0 ...
##  $ Unique.Open.Count      : int  6 1 0 0 5 1 0 7 5 0 ...
##  $ Unique.Click.Count     : int  0 0 0 0 1 0 0 2 0 0 ...
##  $ App.Start              : int  0 12 0 8 38 3 7 13 15 0 ...
##  $ App.Other              : int  0 25 2 9 30 21 9 21 1 3 ...
##  $ App.Completed          : int  2 16 37 6 21 10 9 19 14 8 ...
##  $ App.NULL               : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ App.Onboarding         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ App.Launch.Times       : int  12 27 39 15 63 16 12 25 14 12 ...

df <- limited_version[-c(1, 2, 6:8)]
view(df)
n_obs <- nrow(df)
train_idx <- sample(n_obs * 0.75)
```

```
renewal_train <- df %>% slice(train_idx)
renewal_test <- df %>% slice(-train_idx)

summary(lm(Duration ~ Subscription.Event.Type + Auto.Renew + App.Start +
    App.Other + App.Completed + App.NULL + App.Onboarding + App.Launch.Times
+
    Demo.User + Email.Subscriber, data = limited_version))
##
## Call:
## lm(formula = Duration ~ Subscription.Event.Type + Auto.Renew +
##     App.Start + App.Other + App.Completed + App.NULL + App.Onboarding +
##     App.Launch.Times + Demo.User + Email.Subscriber, data =
limited_version)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -554.77 -113.65  -42.89   83.02 1600.33
##
## Coefficients:
##                          Estimate Std. Error t value
## (Intercept)             208.70576    2.43309  85.778
## Subscription.Event.Type  67.33100    2.03212  33.133
## Auto.Renew               80.02847    1.91243  41.846
## App.Start                -0.01225    0.21143  -0.058
## App.Other                -2.11456    0.20170 -10.484
## App.Completed            -0.30559    0.20192  -1.513
## App.NULL                 -3.05635    2.57617  -1.186
## App.Onboarding           21.50803   45.78211   0.470
## App.Launch.Times          2.14795    0.16175  13.280
## Demo.User               -48.12748    2.17366 -22.141
## Email.Subscriber         16.92739    2.23055   7.589
##                                     Pr(>|t|)
## (Intercept)             < 0.0000000000000002 ***
## Subscription.Event.Type < 0.0000000000000002 ***
## Auto.Renew              < 0.0000000000000002 ***
## App.Start                              0.954
## App.Other               < 0.0000000000000002 ***
## App.Completed                          0.130
## App.NULL                               0.235
## App.Onboarding                         0.639
## App.Launch.Times        < 0.0000000000000002 ***
## Demo.User               < 0.0000000000000002 ***
## Email.Subscriber          0.0000000000000331 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 164.9 on 32055 degrees of freedom
## Multiple R-squared:  0.1181, Adjusted R-squared:  0.1178
## F-statistic: 429.2 on 10 and 32055 DF,  p-value: < 0.00000000000000022
```

```
probability_fit <- lm(renewal_train$Subscription.Event.Type ~ . - Country,
    data = renewal_train)

preds_train <- data.frame(preds = predict(probability_fit, newdata =
renewal_train,
    type = "response"), points = renewal_train$Subscription.Event.Type)

R2(preds_train$preds, renewal_train$Subscription.Event.Type)
## [1] 0.2028974
RMSE(preds_train$preds, renewal_train$Subscription.Event.Type)
## [1] 0.4157709
MAE(preds_train$preds, renewal_train$Subscription.Event.Type)
## [1] 0.3509672

summary(lm(Subscription.Event.Type ~ . - Country, data = df))
##
## Call:
## lm(formula = Subscription.Event.Type ~ . - Country, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1209 -0.2898 -0.1281  0.2633  1.3743
##
## Coefficients: (1 not defined because of singularities)
##                            Estimate  Std. Error t value
## (Intercept)              0.49415365  0.01563409  31.607
## Subscription.Type                NA          NA      NA
## Purchase.Store          -0.22408379  0.00677282 -33.086
## Duration                 0.00052152  0.00001397  37.335
## Demo.User               -0.16439732  0.00973590 -16.886
## Free.Trial.User          0.17129945  0.00896324  19.111
## Auto.Renew               0.06648904  0.00504037  13.191
## User.Type                0.18933282  0.02237770   8.461
## Lead.Platform..App...0. -0.16649623  0.00964170 -17.268
## Email.Subscriber         0.02649558  0.00636035   4.166
## Push.Notifications      -0.29687652  0.02520163 -11.780
## Send.Count              -0.00057046  0.00007574  -7.532
## Open.Count               0.00333952  0.00042837   7.796
## Click.Count             -0.00336999  0.00052747  -6.389
## Unique.Open.Count       -0.00404010  0.00064652  -6.249
## Unique.Click.Count      -0.02274254  0.00399011  -5.700
## App.Start               -0.00306740  0.00053002  -5.787
## App.Other                0.00152054  0.00050561   3.007
## App.Completed           -0.00020055  0.00050447  -0.398
## App.NULL                -0.02342713  0.01088061  -2.153
## App.Onboarding           0.08479697  0.11443114   0.741
## App.Launch.Times         0.00209865  0.00040543   5.176
##                                     Pr(>|t|)
## (Intercept)             < 0.0000000000000002 ***
## Subscription.Type                         NA
## Purchase.Store           < 0.0000000000000002 ***
## Duration                 < 0.0000000000000002 ***
## Demo.User                < 0.0000000000000002 ***
## Free.Trial.User          < 0.0000000000000002 ***
## Auto.Renew               < 0.0000000000000002 ***
## User.Type                < 0.0000000000000002 ***
## Lead.Platform..App...0. < 0.0000000000000002 ***
## Email.Subscriber          0.00003111531086647 ***
## Push.Notifications       < 0.0000000000000002 ***
## Send.Count                0.00000000000005135 ***
## Open.Count                0.00000000000000659 ***
## Click.Count               0.00000000016926000 ***
## Unique.Open.Count         0.00000000041830199 ***
## Unique.Click.Count        0.00000001210536952 ***
## App.Start                 0.00000000721756813 ***
## App.Other                             0.00264 **
## App.Completed                         0.69096
## App.NULL                              0.03132 *
## App.Onboarding                        0.45868
## App.Launch.Times          0.00000022767055642 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4119 on 32045 degrees of freedom
## Multiple R-squared:  0.2005, Adjusted R-squared:    0.2
## F-statistic: 401.7 on 20 and 32045 DF,  p-value: < 0.00000000000000022
```

```
bkwd_fit <- regsubsets(df$Duration ~ . - Country, data = df, method =
"backward",
    nvmax = 8)
## Reordering variables and trying again:
summary(bkwd_fit)
## Subset selection object
## Call: regsubsets.formula(df$Duration ~ . - Country, data = df, method =
"backward",
##     nvmax = 8)
## 21 Variables  (and intercept)
##                          Forced in Forced out
## Subscription.Event.Type     FALSE      FALSE
## Purchase.Store              FALSE      FALSE
## Demo.User                   FALSE      FALSE
## Free.Trial.User             FALSE      FALSE
## Auto.Renew                  FALSE      FALSE
## User.Type                   FALSE      FALSE
## Lead.Platform..App...0.     FALSE      FALSE
## Email.Subscriber            FALSE      FALSE
## Push.Notifications          FALSE      FALSE
## Send.Count                  FALSE      FALSE
## Open.Count                  FALSE      FALSE
## Click.Count                 FALSE      FALSE
## Unique.Open.Count           FALSE      FALSE
## Unique.Click.Count          FALSE      FALSE
## App.Start                   FALSE      FALSE
## App.Other                   FALSE      FALSE
## App.Completed               FALSE      FALSE
## App.NULL                    FALSE      FALSE
## App.Onboarding              FALSE      FALSE
## App.Launch.Times            FALSE      FALSE
## Subscription.Type           FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: backward
##          Subscription.Type Subscription.Event.Type Purchase.Store
## 1  ( 1 ) " "               " "                     " "           
## 2  ( 1 ) " "               "*"                     " "           
## 3  ( 1 ) " "               "*"                     " "           
## 4  ( 1 ) " "               "*"                     " "           
## 5  ( 1 ) " "               "*"                     " "           
## 6  ( 1 ) " "               "*"                     " "           
## 7  ( 1 ) " "               "*"                     "*"           
## 8  ( 1 ) " "               "*"                     "*"           
## 9  ( 1 ) " "               "*"                     "*"           
##          Demo.User Free.Trial.User Auto.Renew User.Type
## 1  ( 1 ) " "       " "             "*"        " "      
## 2  ( 1 ) " "       " "             "*"        " "      
## 3  ( 1 ) "*"       " "             "*"        " "      
## 4  ( 1 ) "*"       "*"             "*"        " "      
## 5  ( 1 ) "*"       "*"             "*"        " "      
## 6  ( 1 ) "*"       "*"             "*"        "*"      
## 7  ( 1 ) "*"       "*"             "*"        "*"      
## 8  ( 1 ) "*"       "*"             "*"        "*"      
## 9  ( 1 ) "*"       "*"             "*"        "*"      
##          Lead.Platform..App...0. Email.Subscriber Push.Notifications
## 1  ( 1 ) " "                     " "              " "               
## 2  ( 1 ) " "                     " "              " "               
## 3  ( 1 ) " "                     " "              " "               
## 4  ( 1 ) " "                     " "              " "               
## 5  ( 1 ) " "                     " "              "*"               
## 6  ( 1 ) " "                     " "              "*"               
## 7  ( 1 ) " "                     " "              "*"               
## 8  ( 1 ) "*"                     " "              "*"               
## 9  ( 1 ) "*"                     " "              "*"               
##          Send.Count Open.Count Click.Count Unique.Open.Count
## 1  ( 1 ) " "        " "        " "         " "              
## 2  ( 1 ) " "        " "        " "         " "              
## 3  ( 1 ) " "        " "        " "         " "              
## 4  ( 1 ) " "        " "        " "         " "              
## 5  ( 1 ) " "        " "        " "         " "              
## 6  ( 1 ) " "        " "        " "         " "              
## 7  ( 1 ) " "        " "        " "         " "              
## 8  ( 1 ) " "        " "        " "         " "              
## 9  ( 1 ) " "        " "        " "         " "
##           Unique.Click.Count App.Start App.Other App.Completed
## 1  ( 1 ) " "                " "       " "       " "
## 2  ( 1 ) " "                " "       " "       " "
## 3  ( 1 ) " "                " "       " "       " "
## 4  ( 1 ) " "                " "       " "       " "
## 5  ( 1 ) " "                " "       " "       " "
## 6  ( 1 ) " "                " "       " "       " "
## 7  ( 1 ) " "                " "       " "       " "
## 8  ( 1 ) " "                " "       " "       " "
## 9  ( 1 ) " "                " "       " "       " "
##           App.NULL App.Onboarding App.Launch.Times
## 1  ( 1 ) " "      " "            " "
## 2  ( 1 ) " "      " "            " "
## 3  ( 1 ) " "      " "            " "
## 4  ( 1 ) " "      " "            " "
## 5  ( 1 ) " "      " "            " "
## 6  ( 1 ) " "      " "            " "
## 7  ( 1 ) " "      " "            " "
## 8  ( 1 ) " "      " "            " "
## 9  ( 1 ) " "      " "            "*"

chisquared <- chisq.test(df$Demo.User, df$Subscription.Event.Type)
chisquared
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$Demo.User and df$Subscription.Event.Type
## X-squared = 4.6716, df = 1, p-value = 0.03066

renewal <- df[df$Subscription.Event.Type == 1, ]
view(renewal)
non_renewal <- df[df$Subscription.Event.Type == 0, ]
view(non_renewal)

summary(lm(Duration ~ Auto.Renew + App.Start + App.Other + App.Completed +
    App.NULL + App.Onboarding + App.Launch.Times + Demo.User +
Email.Subscriber,
    data = renewal))
##
## Call:
## lm(formula = Duration ~ Auto.Renew + App.Start + App.Other +
##     App.Completed + App.NULL + App.Onboarding + App.Launch.Times +
##     Demo.User + Email.Subscriber, data = renewal)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -377.37  -87.76   -1.90   76.62  541.69
##
## Coefficients:
```

```
##                  Estimate Std. Error t value             Pr(>|t|)    
## (Intercept)      283.5280     3.3122  85.601 < 0.0000000000000002 ***
## Auto.Renew        60.2081     2.5144  23.945 < 0.0000000000000002 ***
## App.Start         -0.5114     0.2713  -1.885               0.0594 .  
## App.Other         -1.5594     0.2587  -6.027        0.00000000173 ***
## App.Completed      0.3267     0.2414   1.354               0.1759    
## App.NULL          -0.4496     3.5527  -0.127               0.8993    
## App.Onboarding    17.8641    54.9899   0.325               0.7453    
## App.Launch.Times   0.9278     0.1979   4.689        0.00000278844 ***
## Demo.User        -42.5489     2.8813 -14.767 < 0.0000000000000002 ***
## Email.Subscriber  28.5377     2.9860   9.557 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 122.3 on 9777 degrees of freedom
## Multiple R-squared:  0.0901, Adjusted R-squared:  0.08926 
## F-statistic: 107.6 on 9 and 9777 DF,  p-value: < 0.00000000000000022
```

```
summary(lm(Duration ~ Auto.Renew + App.Start + App.Other + App.Completed +
    App.NULL + App.Onboarding + App.Launch.Times + Demo.User +
Email.Subscriber,
    data = non_renewal))
```

```
## 
## Call:
## lm(formula = Duration ~ Auto.Renew + App.Start + App.Other + 
##     App.Completed + App.NULL + App.Onboarding + App.Launch.Times + 
##     Demo.User + Email.Subscriber, data = non_renewal)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -691.99 -112.84  -67.54   89.39 1609.86 
## 
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      203.19768    3.09838  65.582 < 0.0000000000000002
## Auto.Renew        89.48847    2.53022  35.368 < 0.0000000000000002
## App.Start          0.09845    0.28564   0.345                0.730
## App.Other         -2.18518    0.27332  -7.995   0.0000000000000136
## App.Completed     -0.37245    0.28403  -1.311                0.190
## App.NULL          -1.10379    3.34897  -0.330                0.742
## App.Onboarding    45.01577   63.67964   0.707                0.480
## App.Launch.Times   2.84792    0.22353  12.741 < 0.0000000000000002
## Demo.User        -49.86746    2.87225 -17.362 < 0.0000000000000002
## Email.Subscriber  12.27820    2.92751   4.194  0.00002750367482465
##                     
## (Intercept)      ***
## Auto.Renew       ***
## App.Start           
## App.Other        ***
## App.Completed       
## App.NULL
## App.Onboarding
## App.Launch.Times ***
## Demo.User        ***
## Email.Subscriber ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 179.9 on 22269 degrees of freedom
## Multiple R-squared:  0.07742,    Adjusted R-squared:  0.07705
## F-statistic: 207.7 on 9 and 22269 DF,  p-value: < 0.00000000000000022
```