Churn at QWE Inc

Kelly E. Cronin

9/15/2019

Customer Churn at QWE, Inc.

This analysis looks at predicting customer churn at QWE, Inc.

Preparation

Data Preparation

Load, prepare, explore, and analyze the QWE, Inc. data

```
# Read data
qwe_orig <- read_excel("HBS Case- Predicting Customer Churn at QWE Inc.xlsx",sheet=2)#import data
```

Initial Exploration

```
glimpse(qwe_orig)
```

```
## Observations: 6,347
## Variables: 13
## $ ID                          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11,...
## $ `Customer Age (in months)`  <dbl> 67, 67, 55, 63, 57, 58, 57, 46, 56...
## $ `Churn (1 = Yes, 0 = No)`   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `CHI Score Month 0`         <dbl> 0, 62, 0, 231, 43, 138, 180, 116, ...
## $ `CHI Score 0-1`             <dbl> 0, 4, 0, 1, -1, -10, -5, -11, -7, ...
## $ `Support Cases Month 0`     <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0...
## $ `Support Cases 0-1`         <dbl> 0, 0, 0, -1, 0, 0, 1, 0, -2, 0, 0,...
## $ `SP Month 0`                <dbl> 0, 0, 0, 3, 0, 0, 3, 0, 3, 0, 0, 0...
## $ `SP 0-1`                    <dbl> 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0...
## $ `Logins 0-1`                <dbl> 0, 0, 0, 167, 0, 43, 13, 0, -9, -7...
## $ `Blog Articles 0-1`         <dbl> 0, 0, 0, -8, 0, 0, -1, 0, 1, 0, 3,...
## $ `Views 0-1`                 <dbl> 0, -16, 0, 21996, 9, -33, 907, 38,...
## $ `Days Since Last Login 0-1` <dbl> 31, 31, 31, 0, 31, 0, 0, 6, 7, 14,...
```

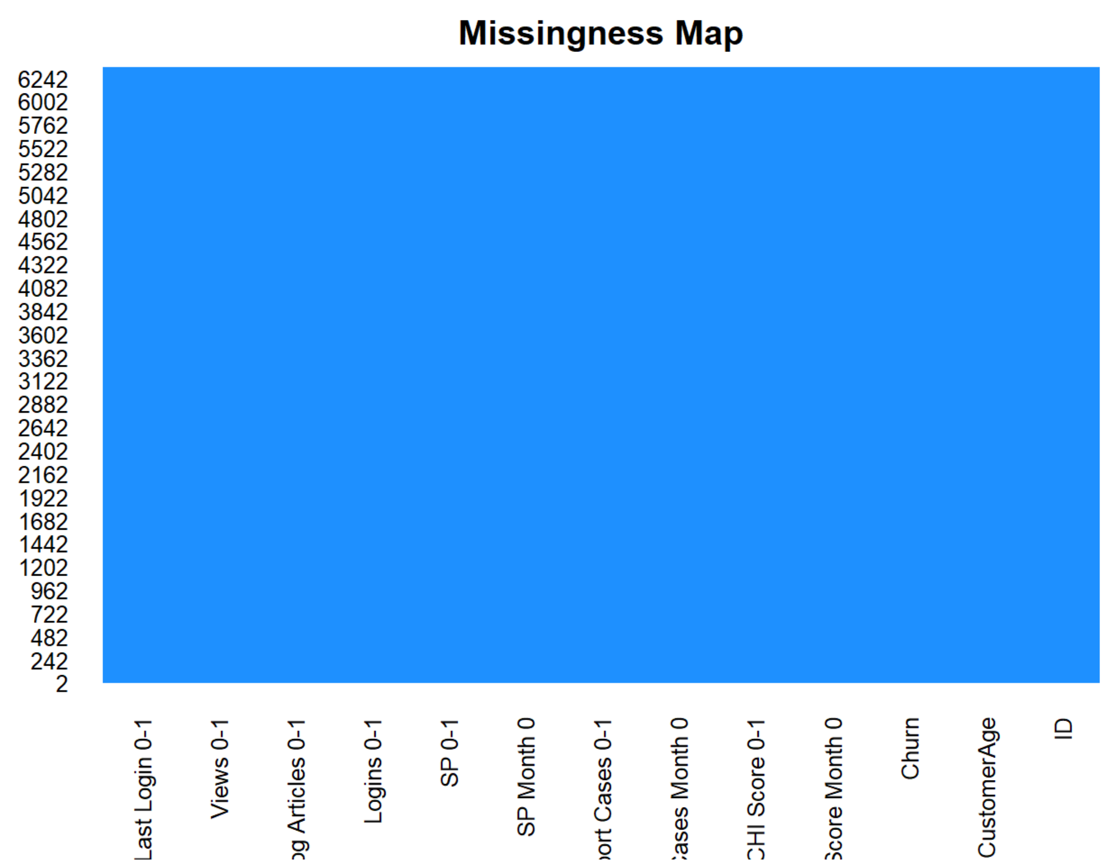
```
qwe<- qwe_orig
colnames(qwe)[colnames(qwe)=="Customer Age (in months)"] <- "CustomerAge"
colnames(qwe)[colnames(qwe)=="Churn (1 = Yes, 0 = No)"] <- "Churn"
qwe<- qwe %>%
  mutate(
    ID = as.character(ID),
    Churn = as.factor(Churn),
  )

summary(qwe)
```

```
##       ID             CustomerAge   Churn    CHI Score Month 0
##  Length:6347        Min.   : 0.0   0:6024   Min.   :  0.00   
##  Class :character   1st Qu.: 5.0   1: 323   1st Qu.: 24.50   
##  Mode  :character   Median :11.0            Median : 87.00   
##                     Mean   :13.9            Mean   : 87.32   
##                     3rd Qu.:20.0            3rd Qu.:139.00   
##                     Max.   :67.0            Max.   :298.00   
##  CHI Score 0-1      Support Cases Month 0 Support Cases 0-1  
##  Min.   :-125.000   Min.   : 0.0000       Min.   :-29.000000  
##  1st Qu.:  -8.000   1st Qu.: 0.0000       1st Qu.:  0.000000  
##  Median :   0.000   Median : 0.0000       Median :  0.000000  
##  Mean   :   5.059   Mean   : 0.7063       Mean   : -0.006932  
##  3rd Qu.:  15.000   3rd Qu.: 1.0000       3rd Qu.:  0.000000  
##  Max.   : 208.000   Max.   :32.0000       Max.   : 31.000000  
##    SP Month 0         SP 0-1            Logins 0-1      Blog Articles 0-1
##  Min.   :0.0000   Min.   :-4.00000   Min.   :-293.00   Min.   :-75.0000  
##  1st Qu.:0.0000   1st Qu.: 0.00000   1st Qu.:  -1.00   1st Qu.:  0.0000  
##  Median :0.0000   Median : 0.00000   Median :   2.00   Median :  0.0000  
##  Mean   :0.8128   Mean   : 0.03017   Mean   :  15.73   Mean   :  0.1572  
##  3rd Qu.:2.6667   3rd Qu.: 0.00000   3rd Qu.:  23.00   3rd Qu.:  0.0000  
##  Max.   :4.0000   Max.   : 4.00000   Max.   : 865.00   Max.   :217.0000  
##    Views 0-1          Days Since Last Login 0-1
##  Min.   :-28322.00   Min.   :-648.000         
##  1st Qu.:   -11.00   1st Qu.:   0.000         
##  Median :     0.00   Median :   0.000         
##  Mean   :    96.31   Mean   :   1.765         
##  3rd Qu.:    27.00   3rd Qu.:   3.000         
##  Max.   :230414.00   Max.   :  61.000
```

```
glimpse(qwe)
```

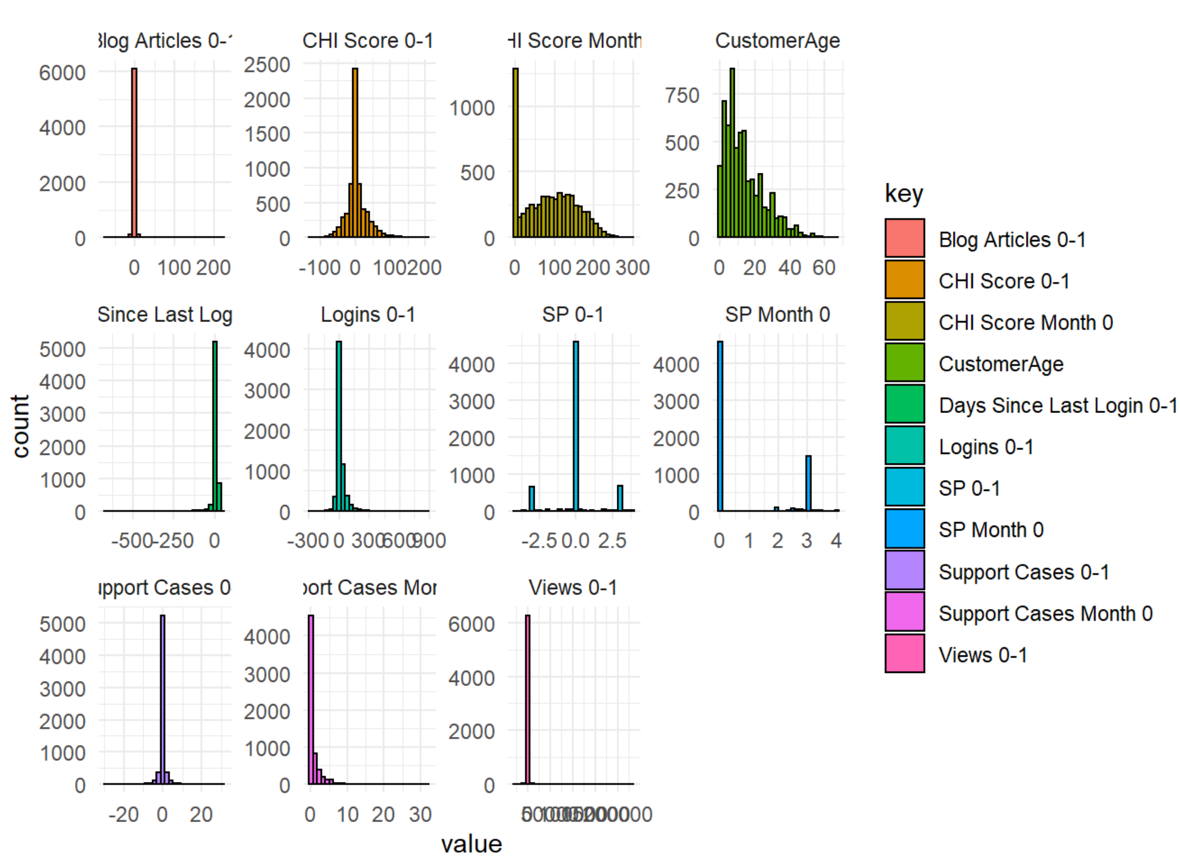
```
## Observations: 6,347
## Variables: 13
## $ ID                          <chr> "1", "2", "3", "4", "5", "6", "7",...
## $ CustomerAge                 <dbl> 67, 67, 55, 63, 57, 58, 57, 46, 56...
## $ Churn                       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ `CHI Score Month 0`         <dbl> 0, 62, 0, 231, 43, 138, 180, 116, ...
## $ `CHI Score 0-1`             <dbl> 0, 4, 0, 1, -1, -10, -5, -11, -7, ...
## $ `Support Cases Month 0`     <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0...
## $ `Support Cases 0-1`         <dbl> 0, 0, 0, -1, 0, 0, 1, 0, -2, 0, 0,...
## $ `SP Month 0`                <dbl> 0, 0, 0, 3, 0, 0, 3, 0, 3, 0, 0, 0...
## $ `SP 0-1`                    <dbl> 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0...
## $ `Logins 0-1`                <dbl> 0, 0, 0, 167, 0, 43, 13, 0, -9, -7...
## $ `Blog Articles 0-1`         <dbl> 0, 0, 0, -8, 0, 0, -1, 0, 1, 0, 3,...
## $ `Views 0-1`                 <dbl> 0, -16, 0, 21996, 9, -33, 907, 38,...
## $ `Days Since Last Login 0-1` <dbl> 31, 31, 31, 0, 31, 0, 0, 6, 7, 14,...
```

```
missmap(qwe, legend=FALSE)
```



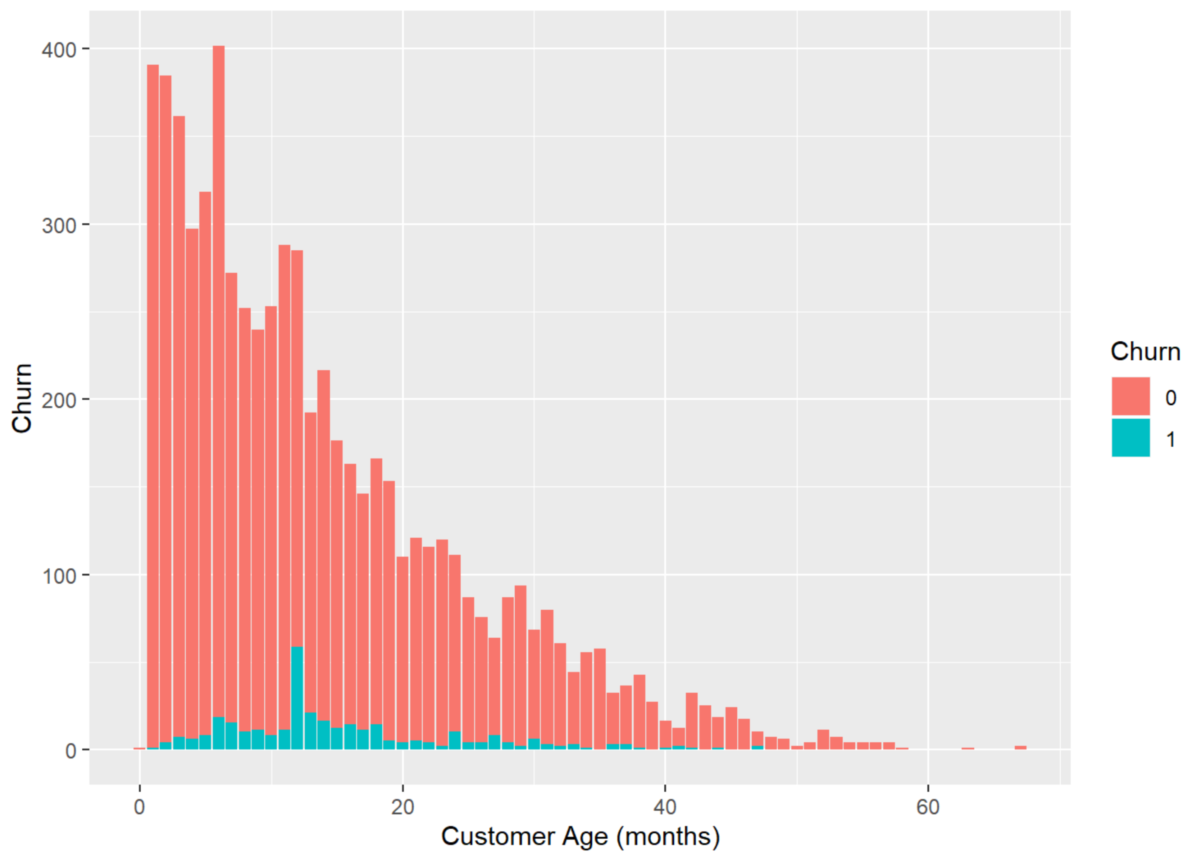

```
qwe %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot() +
  geom_histogram(mapping = aes(x=value,fill=key), color="black") +
  facet_wrap(~ key, scales = "free") +
  theme_minimal()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

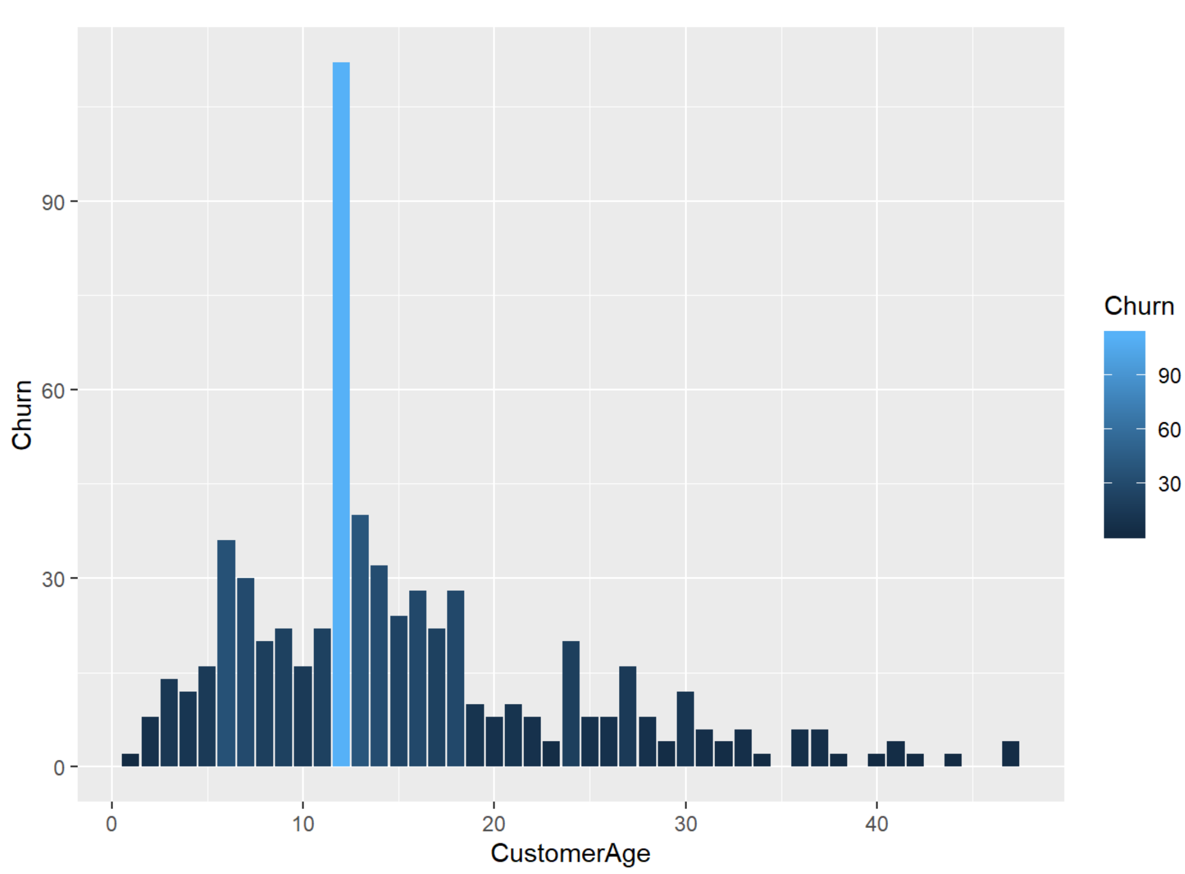


Visualize Churn

```
# Churn rate

#Grouped by Customer Age, what is the average churn?
qwe %>%
  mutate(
    ChurnAvg=mean(as.numeric(Churn))
  ) %>%
  ggplot() +
  geom_col(aes(x=CustomerAge, y=ChurnAvg, fill=Churn))+
  xlab("Customer Age (months)")+
  ylab("Churn")
```




```
# Churn by customer age
#How many people churn at each customer age?
qwe %>%
  filter(Churn==1) %>%
  mutate(Churn=as.numeric(Churn)) %>%
  group_by(CustomerAge) %>%
  summarize(Churn = sum(Churn)) %>%
  ggplot() +
  geom_col(aes(x=CustomerAge, y=Churn, fill=Churn))
```



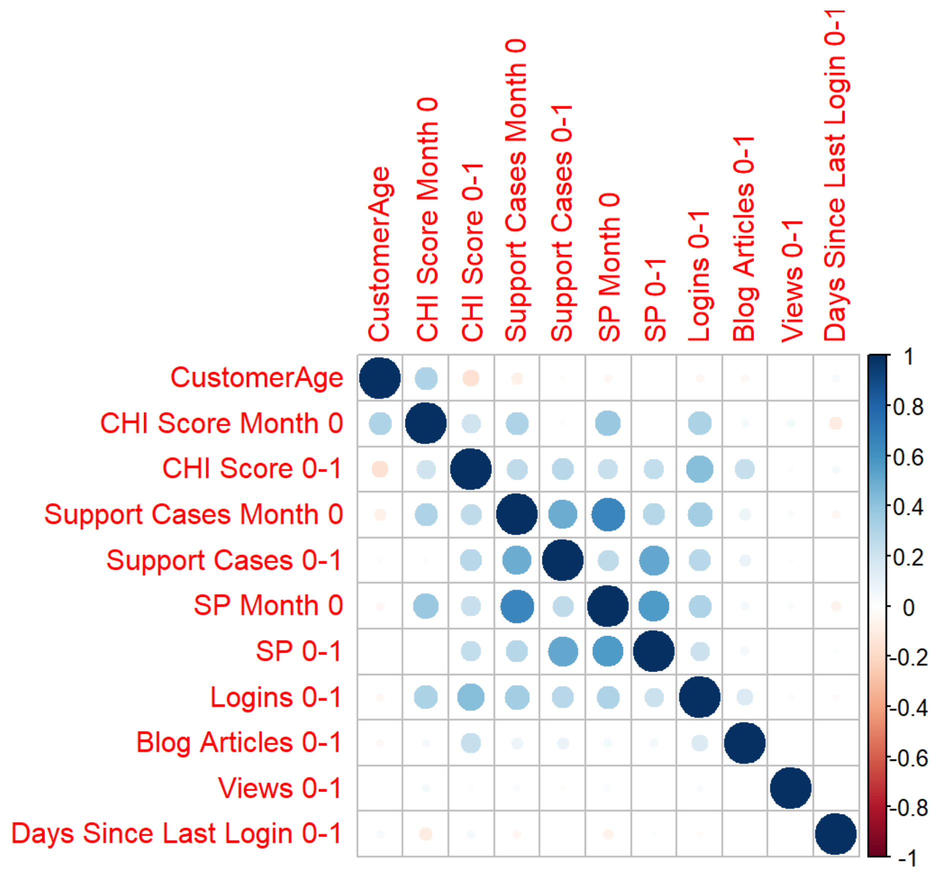

```
 qwe %>%
  filter(Churn==1) %>%
  mutate(Churn=as.numeric(Churn)) %>%
  group_by(CustomerAge) %>%
  summarize(Churn = sum(Churn)) %>%
  arrange(-Churn) %>%
  head(n=1)
```

```
## # A tibble: 1 x 2
##   CustomerAge Churn
##         <dbl> <dbl>
## 1          12   112
```

Wall's intuition about customer age was generally correct: there is a lower instance of churn in customers after 14 months.

Univariate Testing

```
qwe %>%
  keep(is.numeric) %>%
  cor() %>%
  corrplot()
```




```
lapply(c("CustomerAge","`CHI Score Month 0`","`CHI Score 0-1` ","`Support Cases Month 0`","`Support Cases 0-1`","`SP Month 0
` ","`SP 0-1` ","`Logins 0-1` ","`Blog Articles 0-1` ","`Views 0-1` ","`Days Since Last Login 0-1`"),

       function(var) {
           formula    <- as.formula(paste("Churn ~", var))
           res.logist <- glm(formula, data = qwe, family = binomial)
           summary(res.logist)
       })
```

```
## [[1]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4330  -0.3314  -0.3150  -0.3045   2.5010
##
## Coefficients:
##              Estimate Std. Error z value            Pr(>|z|)
## (Intercept) -3.094189   0.093009 -33.268 <0.0000000000000002 ***
## CustomerAge  0.011555   0.004809   2.403              0.0163 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2547.6  on 6345  degrees of freedom
## AIC: 2551.6
##
## Number of Fisher Scoring iterations: 5
##
##
## [[2]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4048  -0.3693  -0.3049  -0.2591   2.7937
##
## Coefficients:
##                       Estimate Std. Error z value             Pr(>|z|)
## (Intercept)         -2.4606438  0.0830552 -29.627 < 0.0000000000000002 ***
## `CHI Score Month 0` -0.0061534  0.0009326  -6.598      0.0000000000417 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2506.6  on 6345  degrees of freedom
## AIC: 2510.6
##
## Number of Fisher Scoring iterations: 6
##
##
## [[3]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6251  -0.3352  -0.3245  -0.2897   2.7386
##
## Coefficients:
##                 Estimate Std. Error z value             Pr(>|z|)
## (Intercept)    -2.917746   0.057648 -50.613 < 0.0000000000000002 ***
## `CHI Score 0-1` -0.011072   0.002068  -5.354         0.0000000859 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2522.8  on 6345  degrees of freedom
## AIC: 2526.8
##
## Number of Fisher Scoring iterations: 6
##
##
## [[4]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3408  -0.3408  -0.3408  -0.3073   3.1473
##
## Coefficients:
##                         Estimate Std. Error z value             Pr(>|z|)
## (Intercept)             -2.81659    0.06135 -45.908 < 0.0000000000000002
## `Support Cases Month 0` -0.21290    0.05867  -3.629             0.000285
##
## (Intercept)             ***
## `Support Cases Month 0` ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2534.4  on 6345  degrees of freedom
## AIC: 2538.4
##
## Number of Fisher Scoring iterations: 6
##
##
## [[5]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3941  -0.3232  -0.3232  -0.3232   2.4815
##
## Coefficients:
##                     Estimate Std. Error z value            Pr(>|z|)
## (Intercept)         -2.92604    0.05712 -51.225 <0.0000000000000002 ***
## `Support Cases 0-1`  0.01322    0.03036   0.435               0.663
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2552.9  on 6345  degrees of freedom
## AIC: 2556.9
##
## Number of Fisher Scoring iterations: 5
##
##
## [[6]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3467  -0.3467  -0.3467  -0.2508   2.7166
##
## Coefficients:
##              Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  -2.78134    0.06292 -44.202 < 0.0000000000000002 ***
## `SP Month 0` -0.22081    0.05128  -4.306            0.0000166 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2532.0  on 6345  degrees of freedom
## AIC: 2536
##
## Number of Fisher Scoring iterations: 6
##
##
## [[7]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3381  -0.3232  -0.3232  -0.3232   2.4674
##
## Coefficients:
##             Estimate Std. Error z value            Pr(>|z|)
## (Intercept) -2.92567    0.05712 -51.224 <0.0000000000000002 ***
## `SP 0-1`    -0.02317    0.03912  -0.592               0.554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2552.8  on 6345  degrees of freedom
## AIC: 2556.8
##
## Number of Fisher Scoring iterations: 5
##
##
## [[8]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7820  -0.3349  -0.3309  -0.3048   3.4480
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept)  -2.852804   0.059094 -48.275 < 0.0000000000000002 ***
## `Logins 0-1` -0.006228   0.001780  -3.499             0.000467 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2539.1  on 6345  degrees of freedom
## AIC: 2543.1
##
## Number of Fisher Scoring iterations: 6
##
##
## [[9]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6431  -0.3233  -0.3233  -0.3202   2.4928
##
## Coefficients:
##                     Estimate Std. Error z value            Pr(>|z|)
## (Intercept)         -2.92543    0.05713 -51.202 <0.0000000000000002 ***
## `Blog Articles 0-1` -0.01939    0.01640  -1.183               0.237
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2551.6  on 6345  degrees of freedom
## AIC: 2555.6
##
## Number of Fisher Scoring iterations: 5
##
##
## [[10]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9790  -0.3233  -0.3232  -0.3225   2.6289
##
## Coefficients:
##                Estimate  Std. Error z value            Pr(>|z|)
## (Intercept) -2.92590125  0.05719363 -51.158 <0.0000000000000002 ***
## `Views 0-1` -0.00008785  0.00003681  -2.387               0.017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2548.2  on 6345  degrees of freedom
## AIC: 2552.2
##
## Number of Fisher Scoring iterations: 5
##
##
## [[11]]
##
## Call:
## glm(formula = formula, family = binomial, data = qwe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5383  -0.3185  -0.3068  -0.3030   3.2665
##
## Coefficients:
##                              Estimate Std. Error z value
## (Intercept)                 -3.032670   0.063289  -47.92
## `Days Since Last Login 0-1`  0.025527   0.004334    5.89
##                                         Pr(>|z|)
## (Intercept)                 < 0.0000000000000002 ***
## `Days Since Last Login 0-1`        0.00000000387 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2553.1  on 6346  degrees of freedom
## Residual deviance: 2517.3  on 6345  degrees of freedom
## AIC: 2521.3
##
## Number of Fisher Scoring iterations: 6
```

List of significant attributes: (Numbers correspond to the results above.) 1. "CustomerAge", 2. " `CHI Score Month 0` ", 3. " `CHI Score 0-1` ", 4. " `Support Cases Month 0` ", 6. " `SP Month 0` ", 8. " `Logins 0-1` ", 10. " `Views 0-1` ", 11. " `Days Since Last Login 0-1` "

List of insignificant attributes:
5. " `Support Cases 0-1` ", 7. " `SP 0-1` ", 9. " `Blog Articles 0-1` "

Logistic Regression

Full Logistic Regression

```
#Set the partitions.
sample_set <- sample(nrow(qwe), round(nrow(qwe)*.75), replace = FALSE)
qwe_train <- qwe[sample_set, ]
qwe_test <- qwe[-sample_set, ]

round(prop.table(table(dplyr::select(qwe, Churn), exclude = NULL)), 4) * 100
```

```
##
##     0     1
## 94.91  5.09
```

```
round(prop.table(table(dplyr::select(qwe_train, Churn), exclude = NULL)), 4) * 100
```

```
##
##     0     1
## 94.71  5.29
```

```
round(prop.table(table(dplyr::select(qwe_test, Churn), exclude = NULL)), 4) * 100
```

```
##
##     0     1
## 95.53  4.47
```

```
#The proportions are roughly equal, so we do not need to further balance them.

logit_mod <-
  speedglm(Churn ~ CustomerAge +`CHI Score Month 0`+`CHI Score 0-1`+`Support Cases Month 0`+`Support Cases 0-1`+`SP Month 0`
+`SP 0-1`+`Logins 0-1`+`Blog Articles 0-1`+`Views 0-1`+`Days Since Last Login 0-1`, family = binomial(), data = qwe_train)
summary(logit_mod)
```

```
## Generalized Linear Model of class 'speedglm':
##
## Call:  speedglm(formula = Churn ~ CustomerAge + `CHI Score Month 0` +      `CHI Score 0-1` + `Support Cases Month 0` + `S
upport Cases 0-1` +      `SP Month 0` + `SP 0-1` + `Logins 0-1` + `Blog Articles 0-1` +      `Views 0-1` + `Days Since Last
Login 0-1`, data = qwe_train,      family = binomial())
##
## Coefficients:
##  ------------------------------------------------------------------
##                              Estimate Std. Error  z value   Pr(>|z|)
## (Intercept)                -2.7378557  1.228e-01 -22.2864 5.010e-110 ***
## CustomerAge                 0.0141165  6.017e-03   2.3460  1.897e-02 *
## `CHI Score Month 0`        -0.0046077  1.378e-03  -3.3434  8.277e-04 ***
## `CHI Score 0-1`            -0.0090319  2.765e-03  -3.2666  1.088e-03 **
## `Support Cases Month 0`    -0.1135462  1.081e-01  -1.0502  2.936e-01
## `Support Cases 0-1`         0.0878473  8.447e-02   1.0400  2.983e-01
## `SP Month 0`                0.0025425  1.131e-01   0.0225  9.821e-01
## `SP 0-1`                    0.0031345  8.465e-02   0.0370  9.705e-01
## `Logins 0-1`                0.0001501  2.278e-03   0.0659  9.475e-01
## `Blog Articles 0-1`         0.0074970  1.925e-02   0.3895  6.969e-01
## `Views 0-1`                -0.0001084  4.326e-05  -2.5053  1.223e-02 *
## `Days Since Last Login 0-1`  0.0160905  4.862e-03   3.3093  9.353e-04 ***
##
## -------------------------------------------------------------------
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ---
## null df: 4759; null deviance: 1971.46;
## residuals df: 4748; residuals deviance: 1892.48;
## # obs.: 4760; # non-zero weighted obs.: 4760;
## AIC: 1916.482; log Likelihood: -946.241;
## RSS: 4778.9; dispersion: 1; iterations: 6;
## rank: 12; max tolerance: 4.73e-09; convergence: TRUE.
```

```
summary(logit_mod)$aic
```

```
## [1] 1916.482
```

Reduced Model

Variables included: * CustomerAge * CHI Score Month 0 * CHI Score 0-1 * Support Cases Month 0 * SP Month 0 * Logins 0-1 * Views 0-1 * Days Since Last Login 0-1

```
logit_reduced <-
  speedglm(Churn ~ CustomerAge +`CHI Score Month 0`+`CHI Score 0-1`+`Support Cases Month 0`++`SP Month 0`+`Logins 0-1`+`View
s 0-1`+`Days Since Last Login 0-1`, family = binomial(), data = qwe_train)
summary(logit_reduced)
```

```
## Generalized Linear Model of class 'speedglm':
## 
## Call:  speedglm(formula = Churn ~ CustomerAge + `CHI Score Month 0` +      `CHI Score 0-1` + `Support Cases Month 0` + +`SP Month 0` +      `Logins 0-1` + `Views 0-1` + `Days Since Last Login 0-1`,      data = qwe_train, family = binomial())
## 
## Coefficients:
##  ------------------------------------------------------------------
##                              Estimate Std. Error  z value   Pr(>|z|)    
## (Intercept)                 -2.7475990  1.221e-01 -22.4947 4.680e-112 ***
## CustomerAge                  0.0154437  5.898e-03   2.6184  8.834e-03 ** 
## `CHI Score Month 0`         -0.0051435  1.305e-03  -3.9413  8.104e-05 ***
## `CHI Score 0-1`             -0.0081647  2.660e-03  -3.0699  2.141e-03 ** 
## `Support Cases Month 0`     -0.0607412  8.625e-02  -0.7042  4.813e-01    
## `SP Month 0`                -0.0006000  8.451e-02  -0.0071  9.943e-01    
## `Logins 0-1`                 0.0006320  2.145e-03   0.2946  7.683e-01    
## `Views 0-1`                 -0.0001131  4.309e-05  -2.6236  8.702e-03 ** 
## `Days Since Last Login 0-1`  0.0159516  4.859e-03   3.2831  1.027e-03 ** 
## 
## ------------------------------------------------------------------
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## ---
## null df: 4759; null deviance: 1971.46;
## residuals df: 4751; residuals deviance: 1894.81;
## # obs.: 4760; # non-zero weighted obs.: 4760;
## AIC: 1912.811; log Likelihood: -947.4055;
## RSS: 4803.7; dispersion: 1; iterations: 6;
## rank: 9; max tolerance: 2.38e-09; convergence: TRUE.
```

```
summary(logit_reduced)$aic
```

```
## [1] 1912.811
```

Yes, the AIC of the reduced model is what I expected since it is lower than the AIC of the full model.

Customer 399

First, establish the predictive model and its cutoff.

```
logit_pred <- predict(logit_mod, qwe_test, type = 'response')

ideal_cutoff <-
  optimalCutoff(
    actuals = qwe_test$Churn,
    predictedScores = logit_pred,
    optimiseFor = "Both"
  )

ideal_cutoff
```

```
## [1] 0.06806327
```

What did Customer 399 actually do? They stayed.

```
Customer_399<- qwe %>%
  filter(ID == "399")

Customer_399$Churn
```

```
## [1] 0
## Levels: 0 1
```

What does the model predict that they would do?

```
logit_pred_399 <- predict(logit_mod, Customer_399, type = 'response')
logit_pred_399
```

```
##          1 
## 0.01897319
```

```
logit_pred_399_result <- ifelse(logit_pred_399 > ideal_cutoff, 1, 0)
logit_pred_399_result
```

```
## 1 
## 0
```

The model predicts that Customer 399 will not leave. Their likelihood of churning is quite low at 1.897%, and that falls below the cutoff point of 6.806%.

Customer 701

What did Customer 701 actually do? They stayed.

```
# 701
Customer_701<- qwe %>%
  filter(ID == "701")

Customer_701$Churn
```

```
## [1] 0
## Levels: 0 1
```

What does the model predict that they would do?

```
logit_pred_701 <- predict(logit_mod, Customer_701, type = 'response')
logit_pred_701
```

```
##          1
## 0.04052075
```

```
logit_pred_701_result <- ifelse(logit_pred_701 > ideal_cutoff, 1, 0)
logit_pred_701_result
```

```
## 1
## 0
```

The model predicts that Customer 701 will not leave. Their likelihood of churning is higher than Customer 399, but still low at 4.052%, and that falls below the cutoff point of 6.806%.

Customer 5020

What did Customer 5020 actually do? They stayed.

```
# 701
Customer_5020<- qwe %>%
  filter(ID == "5020")

Customer_5020$Churn
```

```
## [1] 0
## Levels: 0 1
```

What does the model predict that they would do?

```
logit_pred_5020 <- predict(logit_mod, Customer_5020, type = 'response')
logit_pred_5020
```

```
##          1
## 0.01354933
```

```
logit_pred_5020_result <- ifelse(logit_pred_5020 > ideal_cutoff, 1, 0)
logit_pred_5020_result
```

```
## 1
## 0
```

The model predicts that Customer 5020 will not leave. Their likelihood of churning is quite low at 1.354%, and that falls below the cutoff point of 6.806%.

Segment the Data

Age:
* 0 to 6 months – they were a toss-up, * 6 to 14 months – they were at particular risk of leaving, * 14 or more months – they are less likely to leave.

CHI: * High CHI scores will not likely leave * Low CHI scores or scores that have dropped recently might leave

Service: * If has needed a lot of service or needed service for a serious issue (high SP), may just drop

Log ins: * Large number of log-ins, then less likely to leave.

Blogs: * If they write blogs, then less likely to leave

Views: * More Views, less likely to leave

```
qwe_train2 <- qwe_train %>%
  filter(
    between(CustomerAge,6,14),
    `CHI Score Month 0`<=250,
    `CHI Score 0-1` < 0,
    `Support Cases Month 0`<=15,
    `Support Cases 0-1`< 10,
    `SP Month 0`>=1,
    `SP 0-1` != 0,
    `Logins 0-1` < 100,
    `Blog Articles 0-1` < 10,
    `Views 0-1`< 100,
    `Days Since Last Login 0-1` > -4
  )
qwe_test2 <- anti_join(qwe, qwe_train2)
```

```
## Joining, by = c("ID", "CustomerAge", "Churn", "CHI Score Month 0", "CHI Score 0-1", "Support Cases Month 0", "Support Cases 0-1", "SP Month 0", "SP 0-1", "Logins 0-1", "Blog Articles 0-1", "Views 0-1", "Days Since Last Login 0-1")
```

Logistic Regression

```
logit_mod2 <-
  speedglm(Churn ~ CustomerAge +
             `CHI Score Month 0`+
             `CHI Score 0-1`+
             `Support Cases Month 0`+
             `Support Cases 0-1`+
             `SP Month 0`+
             `SP 0-1`+
             `Logins 0-1`+
             `Blog Articles 0-1`+
             `Views 0-1`+
             `Days Since Last Login 0-1`,
    family = binomial(link = 'logit'),
    data = qwe_train2
  )
summary(logit_mod2)$aic
```

```
## [1] 37.90153
```

```
logit_reduced2 <-
  speedglm(Churn ~ CustomerAge +
             `CHI Score Month 0`+
             `CHI Score 0-1`+
             `Support Cases Month 0`+
             `SP Month 0`+
             `Logins 0-1`+
             `Views 0-1`+
             `Days Since Last Login 0-1`, family = binomial(), data = qwe_train2)
summary(logit_reduced2)
```

```
## Generalized Linear Model of class 'speedglm':
##
## Call:  speedglm(formula = Churn ~ CustomerAge + `CHI Score Month 0` +      `CHI Score 0-1` + `Support Cases Month 0` + `SP Month 0` +      `Logins 0-1` + `Views 0-1` + `Days Since Last Login 0-1`,      data = qwe_train2, family = binomial())
##
## Coefficients:
##  ------------------------------------------------------------------
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.892168   8.001397  0.2365    0.813
## CustomerAge                  0.333333   0.322423  1.0338    0.301
## `CHI Score Month 0`         -0.005319   0.023716 -0.2243    0.823
## `CHI Score 0-1`              0.044244   0.052778  0.8383    0.402
## `Support Cases Month 0`      0.236108   0.592363  0.3986    0.690
## `SP Month 0`                -2.631634   2.164883 -1.2156    0.224
## `Logins 0-1`                 0.026792   0.029178  0.9182    0.358
## `Views 0-1`                  0.004031   0.005126  0.7863    0.432
## `Days Since Last Login 0-1` -0.129863   0.397889 -0.3264    0.744
##
## ------------------------------------------------------------------
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ---
## null df: 44; null deviance: 27;
## residuals df: 36; residuals deviance: 16.98;
## # obs.: 45; # non-zero weighted obs.: 45;
## AIC: 34.98396; log Likelihood: -8.491978;
## RSS: 25.7; dispersion: 1; iterations: 9;
## rank: 9; max tolerance: 5.76e-14; convergence: TRUE.
```

```
summary(logit_reduced2)$aic
```

```
## [1] 34.98396
```

The AIC still went down with the reduced model.

Customer 399 in context

First, establish the predictive model and its cutoff.

```
logit_pred2 <- predict(logit_mod2, qwe_test2, type = 'response')

ideal_cutoff2 <-
  optimalCutoff(
    actuals = qwe_test2$Churn,
    predictedScores = logit_pred2,
    optimiseFor = "Both",
  )

ideal_cutoff2
```

```
## [1] 0.99
```

```
#399
logit_pred_399b <- predict(logit_mod2, Customer_399, type = 'response')
logit_pred_399b
```

```
##         1
## 0.9999068
```

```
logit_pred_399b_result <- ifelse(logit_pred_399b > ideal_cutoff2, 1, 0)
logit_pred_399b_result
```

```
## 1
## 1
```

The new model predicts that Customer 399 will leave with high likelihood.

Customer 701

```
# 701
logit_pred_701b <- predict(logit_mod2, Customer_701, type = 'response')
logit_pred_701b
```

```
##         1
## 0.9994327
```

```
logit_pred_701b_result <- ifelse(logit_pred_701b > ideal_cutoff2, 1, 0)
logit_pred_701b_result
```

```
## 1
## 1
```

The new model predicts that Customer 701 will leave with high likelihood.

Customer 5020

```
# 5020
logit_pred_5020b <- predict(logit_mod2, Customer_5020, type = 'response')
logit_pred_5020b
```

```
## 1
## 1
```

```
logit_pred_5020b_result <- ifelse(logit_pred_5020b > ideal_cutoff2, 1, 0)
logit_pred_5020b_result
```

```
## 1
## 1
```

The model predicts that Customer 5020 will leave with high likelihood.

Model 2 is far too over-fitted to be of use for future predictions, and therefore resulted in false positives when testing on existing data. Though Wall's intuitions anecdotally and, in some cases, individually make sense, there is not enough data to vertically combine each subset – when I attempted this, I ended up back with all 6347 observations – and when they are combined inclusively then the testing partition has too few observations.

Top 10 Lists

```
logit_pred_10 <- predict(logit_mod, qwe, type = 'response')
summary(logit_pred_10)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.03251 0.04792 0.05299 0.06241 0.37784
```

```
qwe_pred<- qwe %>%
  mutate(ChurnPred = logit_pred_10)

top_10<- qwe_pred %>%
  arrange(-ChurnPred) %>%
  head(n=10)
top_10
```

```
## # A tibble: 10 x 14
##    ID    CustomerAge Churn `CHI Score Mont~ `CHI Score 0-1`
##    <chr>       <dbl> <fct>            <dbl>           <dbl>
##  1 2287           34 0                  227               7
##  2 357            12 1                  203              25
##  3 109            40 0                    0            -125
##  4 1971           30 0                    0            -113
##  5 2025           28 0                   18             -15
##  6 1              67 0                    0               0
##  7 2076           29 0                   29             -69
##  8 929            11 0                  123              35
##  9 14             57 0                    0               0
## 10 1363           41 1                    0             -34
## # ... with 9 more variables: `Support Cases Month 0` <dbl>, `Support Cases
## #   0-1` <dbl>, `SP Month 0` <dbl>, `SP 0-1` <dbl>, `Logins 0-1` <dbl>,
## #   `Blog Articles 0-1` <dbl>, `Views 0-1` <dbl>, `Days Since Last Login
## #   0-1` <dbl>, ChurnPred <dbl>
```

```
top_10$ChurnPred
```

```
##  [1] 0.3778369 0.2942663 0.2838471 0.2328919 0.2160127 0.2153039 0.2071560
##  [8] 0.2054781 0.1924123 0.1923218
```

I chose the reduced Model 1 because it did not suffer from over-fitting like Model 2 (full and reduced) did while maintaining a lower AIC than the full Model 1.