- Nha Nguyen
- 671491808

HINTS Analysis Feature Engineering

- Target variable D4
- Access to online medical record

```r
# Convert target variable 'AccessOnlineRecord' column to factor
finaldata$AccessOnlineRecord <- ifelse(finaldata$AccessOnlineRecord == 0, "No","Yes")
finaldata$AccessOnlineRecord <- as.factor(finaldata$AccessOnlineRecord)
```

Target Variable

- **D4 - AccessOnlineRecord**
- How often do you access to the online medical record in the last 12 months? -> Binary Target Variable

- **Approach**

- Divide target variable into 2 classes: “0” and “1”
- Freq = 0 time -> no access to online medical record <-> **Class=0 or Class= “no”**
- Freq >=1 time -> have access to online medical record <-> **Class=1 or Class = “yes”**

HINTS - Feature Selection

Demographics

- P1. Age
- P2. Birth Gender
- P16. Income Ranges
- P5. Occupation Status - Occupation_Employed,Occupation_Homemaker, Occupation_Student,Occupation_Retired, Occupation_Disabled
- P6. Marital Status

Health Status/ Condition/ Practice

- H1. General Health
- C2. Frequency of going to Provider -FreqGoProvider

Technology Usage/Access / Behavioral Pattern

- B5.Electronic means purposes - Electronic_SelfHealthInfo,Electronic_TalkDoctor, Electronic_TestResults, Electronic_MadeAppts
- B7.Access to tablet wellness app - TabletHealthWellnessApps
- B14. Internet Purpose in the past 12 months - IntRsn_VisitedSocNet, IntRsn_SharedSocNet, IntRsn_SupportGroup,IntRsn_YouTube,

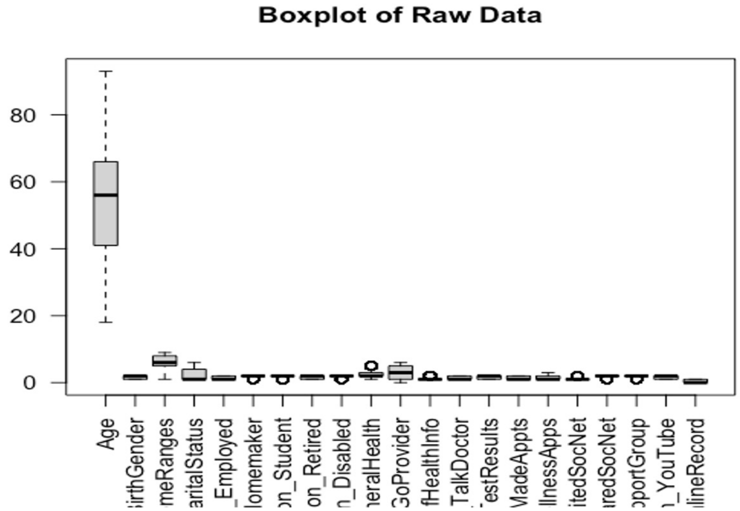
I. Data Cleaning

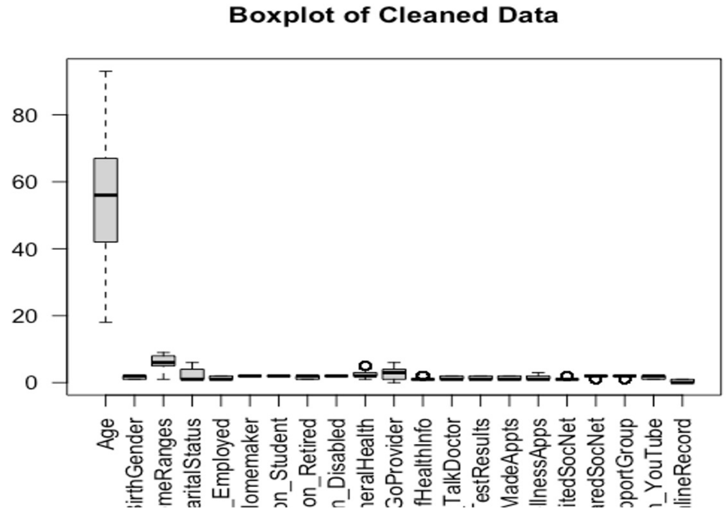
- Remove negative value in the target variable column and any other columns by replacing those values with NA values. Then, omit NA rows.
- 3865 observations drop to 2749 observations

```
#Remove negative value in the Target Variable column "AccessOnlineRecord" and any other column
df<-data.frame(rawdata)
df[df<0] <- NA

#Omit NA rows
df1 <- na.omit(df)
df1
```

I. Data Cleaning (cont.)

- Check for outliers using Boxplot and Z-score
- z score tells how many standard deviations a given value is from the mean. We define an observation to be an outlier if it has a z-score less than -3 or greater than 3.
- remove rows that have at least one z-score with an absolute value greater than 3.



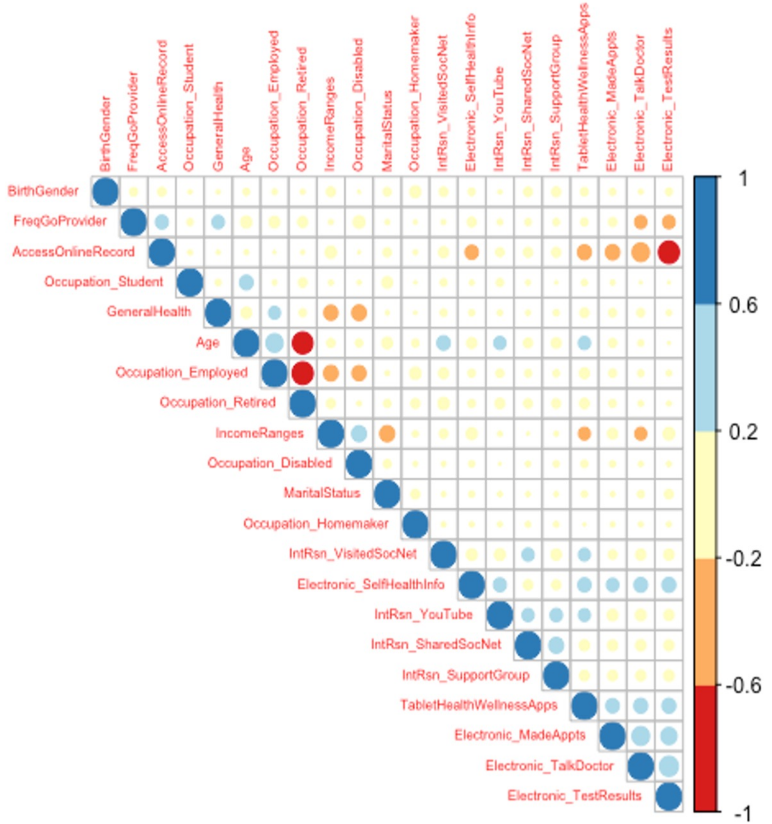


- Remove 454 outliers

II. Exploratory Data Analysis

```
#Check for Correlation
corr_matrix <- round(cor(df1), digits = 2)
corrplot(corr_matrix, type = "upper",order = "hclust",col=brewer.pal(n=5, name= "RdYlBu"),tl.cex=0.5)
```

- As seen in the heatmap, "**Electronic_Test Result"** is the most correlated with the target variable.
- Variables that have moderate correlation with the target variable are **Electronic_TalkDoctor, Electronic_MadeAppts, TabletHealthWellnessApps and Electronic_SelfHealthInfo**



III. Deploy ML Model - Logistic Regression Model

Model 1

```
lm1 <- glm(`AccessOnlineRecord`~ .,
           data= train,family="binomial")
summary(lm1)
```

```
Call:
glm(formula = AccessOnlineRecord ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3769  -0.5768  -0.2801   0.6217   2.3992

Coefficients: (3 not defined because of singularities)
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                4.8534074  1.1604820   4.182 2.89e-05 ***
Age                        0.0002124  0.0057250   0.037 0.970411
BirthGender                0.4468657  0.1340201   3.334 0.000855 ***
IncomeRanges               0.0735575  0.0387155   1.900 0.057440 .
MaritalStatus             -0.0235330  0.0372946  -0.631 0.528039
Occupation_Employed       -0.1499063  0.2609683  -0.574 0.565681
Occupation_Homemaker              NA         NA      NA       NA
Occupation_Student                NA         NA      NA       NA
Occupation_Retired        -0.3959766  0.2905937  -1.363 0.172994
Occupation_Disabled               NA         NA      NA       NA
GeneralHealth             -0.0654262  0.0792988  -0.825 0.409338
FreqGoProvider             0.1581222  0.0381596   4.144 3.42e-05 ***
Electronic_SelfHealthInfo -0.4188193  0.1894915  -2.210 0.027089 *
Electronic_TalkDoctor     -0.5972696  0.1524182  -3.919 8.91e-05 ***
Electronic_TestResults    -2.6455059  0.1422043 -18.604  < 2e-16 ***
Electronic_MadeAppts      -0.0624035  0.1491201  -0.418 0.675597
TabletHealthWellnessApps  -0.4344160  0.1282810  -3.386 0.000708 ***
IntRsn_VisitedSocNet      -0.3068754  0.1626629  -1.887 0.059218 .
IntRsn_SharedSocNet        0.3082159  0.1988623   1.550 0.121166
IntRsn_SupportGroup        0.0713991  0.2257092   0.316 0.751750
IntRsn_YouTube             0.1145889  0.1404508   0.816 0.414577
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2533.9  on 1828  degrees of freedom
Residual deviance: 1571.0  on 1811  degrees of freedom
AIC: 1607
```

III. Model (cont.)

- Run Anova test
- With 95% confidence level, a variable having $p<0.05$ is considered important predictors.
- From the output, variables such as "Electronic_SelfHealthInfo", "Electronic_TalkDoctor", "Electronic_TestResults", "TabletHealthWellnessApps" should be considered for the second model since they are good predictors.

```
> anova(lm1, test = 'Chisq')
Analysis of Deviance Table

Model: binomial, link: logit

Response: AccessOnlineRecord

Terms added sequentially (first to last)


                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                       1828     2533.9
Age                        1     0.04      1827     2533.8 0.8485911
BirthGender                1    16.78      1826     2517.1 4.187e-05 ***
IncomeRanges               1    80.81      1825     2436.2 < 2.2e-16 ***
MaritalStatus              1     2.50      1824     2433.7 0.1136110
Occupation_Employed        1     4.55      1823     2429.2 0.0330023 *
Occupation_Homemaker       0     0.00      1823     2429.2
Occupation_Student         0     0.00      1823     2429.2
Occupation_Retired         1     1.72      1822     2427.5 0.1890926
Occupation_Disabled        0     0.00      1822     2427.5
GeneralHealth              1     0.01      1821     2427.5 0.9270990
FreqGoProvider             1   103.00      1820     2324.5 < 2.2e-16 ***
Electronic_SelfHealthInfo  1    78.55      1819     2245.9 < 2.2e-16 ***
Electronic_TalkDoctor      1   215.23      1818     2030.7 < 2.2e-16 ***
Electronic_TestResults     1   441.26      1817     1589.4 < 2.2e-16 ***
Electronic_MadeAppts       1     0.27      1816     1589.2 0.6021069
TabletHealthWellnessApps   1    11.57      1815     1577.6 0.0006686 ***
IntRsn_VisitedSocNet       1     2.32      1814     1575.3 0.1280660
IntRsn_SharedSocNet        1     3.46      1813     1571.8 0.0627574 .
IntRsn_SupportGroup        1     0.18      1812     1571.6 0.6700716
IntRsn_YouTube             1     0.67      1811     1571.0 0.4140309
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

III. Deploy ML Model - Logistic Regression Model

Model 2

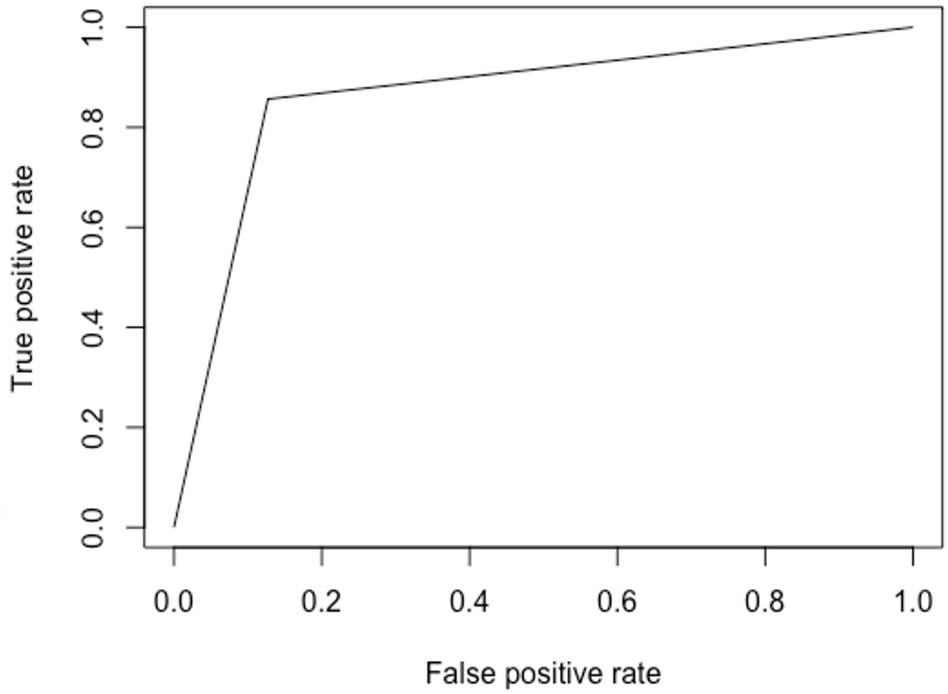
```
> summary(lm2)

Call:
glm(formula = AccessOnlineRecord ~ Electronic_TalkDoctor + Electronic_MadeAppts +
    TabletHealthWellnessApps + Electronic_SelfHealthInfo + Electronic_TestResults,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9725  -0.6081  -0.3975   0.5561   2.2830

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                6.080400   0.312680  19.446  < 2e-16 ***
Electronic_TalkDoctor     -0.683263   0.147906  -4.620 3.85e-06 ***
Electronic_MadeAppts      -0.002647   0.145473  -0.018 0.985481
TabletHealthWellnessApps  -0.467625   0.121497  -3.849 0.000119 ***
Electronic_SelfHealthInfo -0.436668   0.180125  -2.424 0.015340 *
Electronic_TestResults    -2.699184   0.137308 -19.658  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2533.9  on 1828  degrees of freedom
Residual deviance: 1619.6  on 1823  degrees of freedom
AIC: 1631.6
```

IV. Model Comparison

```
> #compare two models
> anova(lm1,lm2,test = "Chisq")
Analysis of Deviance Table

Model 1: AccessOnlineRecord ~ Age + BirthGender + IncomeRanges + MaritalStatus +
    Occupation_Employed + Occupation_Homemaker + Occupation_Student +
    Occupation_Retired + Occupation_Disabled + GeneralHealth +
    FreqGoProvider + Electronic_SelfHealthInfo + Electronic_TalkDoctor +
    Electronic_TestResults + Electronic_MadeAppts + TabletHealthWellnessApps +
    IntRsn_VisitedSocNet + IntRsn_SharedSocNet + IntRsn_SupportGroup +
    IntRsn_YouTube
Model 2: AccessOnlineRecord ~ Electronic_TalkDoctor + Electronic_MadeAppts +
    TabletHealthWellnessApps + Electronic_SelfHealthInfo + Electronic_TestResults
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1      1811     1571.0
2      1823     1619.7 -12  -48.695 2.365e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

V. Evaluate Model

```r
log_predict <- predict(lm2,newdata = test,type = "response")
log_predict <- ifelse(log_predict > 0.5,1,0)

#Plot ROC Curve and Calculate AUC

pr <- prediction(log_predict,test$AccessOnlineRecord)
perf <- performance(pr,measure = "tpr",x.measure = "fpr")
auc(test$AccessOnlineRecord,log_predict) #86.47%
plot(perf)
```



Thank you!