ML Final Project

Sandra

12/18/2021

```
rm(list=ls())
```

```
library(psych)
library(MASS)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(rpart)
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:psych':
##
##     outlier
```

```
library(nnet)
library(e1071)
library(gam)
```

```
## Loading required package: splines

## Loading required package: foreach

## Loaded gam 1.20
```

```
# Data Preparation
labels = read.csv("training_set_labels.csv", header = TRUE)
df = read.csv("training_set_features.csv", header = TRUE)
data = merge(df, labels, by = "respondent_id")
data = data[,-c(12,20:22,28:29,31:32,35:36,38)]
head(data,5)
```

```
##   respondent_id h1n1_concern h1n1_knowledge behavioral_antiviral_meds
## 1             0            1              0                         0
## 2             1            3              2                         0
## 3             2            1              1                         0
## 4             3            1              1                         0
## 5             4            2              1                         0
##   behavioral_avoidance behavioral_face_mask behavioral_wash_hands
## 1                    0                    0                     0
## 2                    1                    0                     1
## 3                    1                    0                     0
## 4                    1                    0                     1
## 5                    1                    0                     1
##   behavioral_large_gatherings behavioral_outside_home behavioral_touch_face
## 1                           0                       1                     1
## 2                           0                       1                     1
## 3                           0                       0                     0
## 4                           1                       0                     0
## 5                           1                       0                     1
##   doctor_recc_h1n1 chronic_med_condition child_under_6_months health_worker
## 1                0                     0                    0             0
## 2                0                     0                    0             0
## 3               NA                     1                    0             0
## 4                0                     1                    0             0
## 5                0                     0                    0             0
##   health_insurance opinion_h1n1_vacc_effective opinion_h1n1_risk
## 1                1                           3                 1
## 2                1                           5                 4
## 3               NA                           3                 1
## 4               NA                           3                 3
## 5               NA                           3                 3
##   opinion_h1n1_sick_from_vacc     age_group        education  race    sex
## 1                           2 55 - 64 Years       < 12 Years White Female
## 2                           4 35 - 44 Years         12 Years White   Male
## 3                           1 18 - 34 Years College Graduate White   Male
## 4                           5     65+ Years         12 Years White Female
## 5                           2 45 - 54 Years     Some College White Female
##              income_poverty  employment_status household_adults
## 1             Below Poverty Not in Labor Force                0
## 2             Below Poverty           Employed                0
## 3 <= $75,000, Above Poverty           Employed                2
## 4             Below Poverty Not in Labor Force                0
## 5 <= $75,000, Above Poverty           Employed                1
##   household_children h1n1_vaccine
## 1                  0            0
## 2                  0            0
## 3                  0            0
## 4                  0            0
## 5                  0            0
```

```
describe(data)
```

```
##                             vars     n     mean      sd median  trimmed     mad
## respondent_id                  1 26707 13353.00 7709.79  13353 13353.00 9899.32
## h1n1_concern                   2 26615     1.62    0.91      2     1.65    1.48
## h1n1_knowledge                 3 26591     1.26    0.62      1     1.32    0.00
## behavioral_antiviral_meds      4 26636     0.05    0.22      0     0.00    0.00
## behavioral_avoidance           5 26499     0.73    0.45      1     0.78    0.00
## behavioral_face_mask           6 26688     0.07    0.25      0     0.00    0.00
## behavioral_wash_hands          7 26665     0.83    0.38      1     0.91    0.00
## behavioral_large_gatherings    8 26620     0.36    0.48      0     0.32    0.00
## behavioral_outside_home        9 26625     0.34    0.47      0     0.30    0.00
## behavioral_touch_face         10 26579     0.68    0.47      1     0.72    0.00
## doctor_recc_h1n1              11 24547     0.22    0.41      0     0.15    0.00
## chronic_med_condition         12 25736     0.28    0.45      0     0.23    0.00
## child_under_6_months          13 25887     0.08    0.28      0     0.00    0.00
## health_worker                 14 25903     0.11    0.32      0     0.01    0.00
## health_insurance              15 14433     0.88    0.33      1     0.97    0.00
## opinion_h1n1_vacc_effective   16 26316     3.85    1.01      4     3.98    1.48
## opinion_h1n1_risk             17 26319     2.34    1.29      2     2.22    1.48
## opinion_h1n1_sick_from_vacc   18 26312     2.36    1.36      2     2.22    1.48
## age_group*                    19 26707     3.19    1.46      3     3.23    1.48
## education*                    20 26707     3.71    1.11      4     3.83    1.48
## race*                         21 26707     3.57    0.92      4     3.81    0.00
## sex*                          22 26707     1.41    0.49      1     1.38    0.00
## income_poverty*               23 26707     2.29    0.86      2     2.24    1.48
## employment_status*            24 26707     2.44    0.68      2     2.42    0.00
## household_adults              25 26458     0.89    0.75      1     0.80    0.00
## household_children            26 26458     0.53    0.93      0     0.34    0.00
## h1n1_vaccine                  27 26707     0.21    0.41      0     0.14    0.00
##                             min   max range  skew kurtosis    se
## respondent_id                 0 26706 26706  0.00    -1.20 47.18
## h1n1_concern                  0     3     3 -0.16    -0.77  0.01
## h1n1_knowledge                0     2     2 -0.24    -0.62  0.00
## behavioral_antiviral_meds     0     1     1  4.19    15.52  0.00
## behavioral_avoidance          0     1     1 -1.01    -0.98  0.00
## behavioral_face_mask          0     1     1  3.40     9.57  0.00
## behavioral_wash_hands         0     1     1 -1.72     0.95  0.00
## behavioral_large_gatherings   0     1     1  0.59    -1.65  0.00
## behavioral_outside_home       0     1     1  0.69    -1.53  0.00
## behavioral_touch_face         0     1     1 -0.76    -1.43  0.00
## doctor_recc_h1n1              0     1     1  1.35    -0.18  0.00
## chronic_med_condition         0     1     1  0.96    -1.07  0.00
## child_under_6_months          0     1     1  3.03     7.20  0.00
## health_worker                 0     1     1  2.46     4.06  0.00
## health_insurance              0     1     1 -2.33     3.45  0.00
## opinion_h1n1_vacc_effective   1     5     4 -0.90     0.51  0.01
## opinion_h1n1_risk             1     5     4  0.67    -0.85  0.01
## opinion_h1n1_sick_from_vacc   1     5     4  0.65    -1.02  0.01
## age_group*                    1     5     4 -0.21    -1.31  0.01
## education*                    1     5     4 -0.74    -0.08  0.01
## race*                         1     4     3 -1.97     2.36  0.01
## sex*                          1     2     1  0.38    -1.85  0.00
## income_poverty*               1     4     3  0.36    -0.46  0.01
## employment_status*            1     4     3  0.22    -0.15  0.00
## household_adults              0     3     3  0.79     0.72  0.00
## household_children            0     3     3  1.54     1.04  0.01
## h1n1_vaccine                  0     1     1  1.41    -0.02  0.00
```

```
# describe(test)
```

```
# Split data into train and test dataset
set.seed(100)
trn <- runif(nrow(data)) < 0.7
```

```
train <- data[trn == TRUE,]
test <- data[trn == FALSE,]
train <- train[complete.cases(train), -c(1)]
test <- test[complete.cases(test), -c(1)]
```

```
# Confusion Matrix
TPR <- function(y,yhat)  { sum(y==1 & yhat==1) / sum(y==1) }
TNR <- function(y,yhat)  { sum(y==0 & yhat==0) / sum(y==0) }
ME <- function(y, yhat)  { (sum(y==1 & yhat==0)+sum(y==0 & yhat==1))/sum(y==1|y==0)}
```

** PART I – LINEAR REGRESSION ******************************* R-square: The larger R-squared, the better the regression model fits your data. The R^2 for improved model m1 and original ols_full are similar to 0.33. The low R^2 presents the model cannot fit the sample data well.

AIC: The lower the value for AIC, the better the fit of the model. The absolute value of the AIC value is not important. It can be positive or negative. When applying OLS regression, AIC is around -26848, which means it is quite nice to fit the model by using “backward selection” method.

```
# OLS regression
ols_full <- lm(h1n1_vaccine ~ ., data = train)
summary(ols_full)
```

```
##
## Call:
## lm(formula = h1n1_vaccine ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04503 -0.23520 -0.08938  0.21068  1.34183
##
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -0.4063816  0.0674994  -6.021 1.80e-09
## h1n1_concern                             0.0032565  0.0051300   0.635 0.525583
## h1n1_knowledge                           0.0327621  0.0069211   4.734 2.24e-06
## behavioral_antiviral_meds                0.0053029  0.0178045   0.298 0.765829
## behavioral_avoidance                    -0.0074204  0.0097222  -0.763 0.445336
## behavioral_face_mask                     0.0063845  0.0159191   0.401 0.688385
## behavioral_wash_hands                   -0.0036640  0.0112445  -0.326 0.744549
## behavioral_large_gatherings             -0.0173149  0.0101737  -1.702 0.088803
## behavioral_outside_home                  0.0003631  0.0103854   0.035 0.972108
## behavioral_touch_face                    0.0064196  0.0094170   0.682 0.495446
## doctor_recc_h1n1                         0.3037252  0.0094170  32.253  < 2e-16
## chronic_med_condition                    0.0205963  0.0088502   2.327 0.019975
## child_under_6_months                     0.0397643  0.0140388   2.832 0.004629
## health_worker                            0.1466924  0.0126316  11.613  < 2e-16
## health_insurance                         0.0689890  0.0128815   5.356 8.72e-08
## opinion_h1n1_vacc_effective              0.0783474  0.0041505  18.877  < 2e-16
## opinion_h1n1_risk                        0.0982310  0.0035213  27.896  < 2e-16
## opinion_h1n1_sick_from_vacc             -0.0053576  0.0031073  -1.724 0.084699
## age_group35 - 44 Years                  -0.0030571  0.0139486  -0.219 0.826525
## age_group45 - 54 Years                  -0.0048162  0.0126637  -0.380 0.703719
## age_group55 - 64 Years                   0.0601063  0.0130509   4.606 4.17e-06
## age_group65+ Years                       0.0806138  0.0141111   5.713 1.14e-08
## education< 12 Years                     -0.1044783  0.0499442  -2.092 0.036474
## education12 Years                       -0.0710011  0.0487569  -1.456 0.145363
## educationCollege Graduate               -0.0330952  0.0486490  -0.680 0.496341
## educationSome College                   -0.0582258  0.0486812  -1.196 0.231701
## raceHispanic                             0.0593123  0.0207074   2.864 0.004188
## raceOther or Multiple                    0.0782352  0.0209497   3.734 0.000189
## raceWhite                                0.0711821  0.0148147   4.805 1.57e-06
## sexMale                                  0.0475699  0.0082346   5.777 7.85e-09
## income_poverty<= $75,000, Above Poverty -0.0375428  0.0128751  -2.916 0.003555
## income_poverty> $75,000                 -0.0210322  0.0147222  -1.429 0.153151
## income_povertyBelow Poverty             -0.0488599  0.0172274  -2.836 0.004575
## employment_statusEmployed               -0.0513660  0.0462998  -1.109 0.267276
## employment_statusNot in Labor Force     -0.0455550  0.0464037  -0.982 0.326267
## employment_statusUnemployed             -0.0436792  0.0485291  -0.900 0.368110
## household_adults                        -0.0008568  0.0052718  -0.163 0.870898
## household_children                       0.0008366  0.0049505   0.169 0.865800
##
## (Intercept)                             ***
## h1n1_concern
## h1n1_knowledge                          ***
## behavioral_antiviral_meds
## behavioral_avoidance
## behavioral_face_mask
## behavioral_wash_hands
## behavioral_large_gatherings             .
## behavioral_outside_home
## behavioral_touch_face
## doctor_recc_h1n1                        ***
## chronic_med_condition                   *
## child_under_6_months                    **
## health_worker                           ***
## health_insurance                        ***
## opinion_h1n1_vacc_effective             ***
## opinion_h1n1_risk                       ***
## opinion_h1n1_sick_from_vacc             .
## age_group35 - 44 Years
## age_group45 - 54 Years
## age_group55 - 64 Years                  ***
## age_group65+ Years                      ***
## education< 12 Years                     *
## education12 Years
## educationCollege Graduate
## educationSome College
## raceHispanic                            **
## raceOther or Multiple                   ***
## raceWhite                               ***
## sexMale                                 ***
## income_poverty<= $75,000, Above Poverty **
## income_poverty> $75,000
## income_povertyBelow Poverty             **
## employment_statusEmployed
## employment_statusNot in Labor Force
## employment_statusUnemployed
## household_adults
## household_children
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3725 on 9515 degrees of freedom
## Multiple R-squared:  0.3352, Adjusted R-squared:  0.3326
## F-statistic: 129.6 on 37 and 9515 DF,  p-value: < 2.2e-16
```

```
summary(ols_full)$adj.r.squared
```

```
## [1] 0.332576
```

```
# Enhance the model
m1 <- lm(h1n1_vaccine ~ h1n1_knowledge+doctor_recc_h1n1+child_under_6_months+
           health_worker+health_insurance+opinion_h1n1_vacc_effective+
           opinion_h1n1_risk+age_group+race+sex, data = train)
summary(m1)
```

```
##
## Call:
## lm(formula = h1n1_vaccine ~ h1n1_knowledge + doctor_recc_h1n1 +
##     child_under_6_months + health_worker + health_insurance +
##     opinion_h1n1_vacc_effective + opinion_h1n1_risk + age_group +
##     race + sex, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07895 -0.23751 -0.08588  0.21171  1.32743
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.5849257  0.0238622 -24.513  < 2e-16 ***
## h1n1_knowledge              0.0439498  0.0066257   6.633 3.46e-11 ***
## doctor_recc_h1n1            0.3054702  0.0093314  32.736  < 2e-16 ***
## child_under_6_months        0.0359599  0.0139931   2.570   0.0102 *
## health_worker               0.1508297  0.0123707  12.192  < 2e-16 ***
## health_insurance            0.0828767  0.0123101   6.732 1.76e-11 ***
## opinion_h1n1_vacc_effective 0.0787409  0.0040401  19.490  < 2e-16 ***
## opinion_h1n1_risk           0.0958808  0.0032802  29.230  < 2e-16 ***
## age_group35 - 44 Years      0.0029464  0.0135605   0.217   0.8280
## age_group45 - 54 Years      0.0002866  0.0123350   0.023   0.9815
## age_group55 - 64 Years      0.0654181  0.0121864   5.368 8.14e-08 ***
## age_group65+ Years          0.0841752  0.0120478   6.987 3.00e-12 ***
## raceHispanic                0.0507913  0.0205238   2.475   0.0134 *
## raceOther or Multiple       0.0825627  0.0208678   3.956 7.66e-05 ***
## raceWhite                   0.0795146  0.0144848   5.490 4.13e-08 ***
## sexMale                     0.0506710  0.0079326   6.388 1.76e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3733 on 9537 degrees of freedom
## Multiple R-squared:  0.3309, Adjusted R-squared:  0.3299
## F-statistic: 314.5 on 15 and 9537 DF,  p-value: < 2.2e-16
```

```
summary(m1)$adj.r.squared
```

```
## [1] 0.3298787
```

```
# Predictions
yhat.ols <- predict(m1, newdata = test)
table(test$h1n1_vaccine, (yhat.ols > 0.5))
```

```
##
##     FALSE TRUE
##   0  2547  199
##   1   554  665
```

```
TPR(test$h1n1_vaccine, (yhat.ols > 0.5))
```

```
## [1] 0.5455291
```

```
TNR(test$h1n1_vaccine, (yhat.ols > 0.5))
```

```
## [1] 0.927531
```

```
ME(test$h1n1_vaccine, (yhat.ols > 0.5))
```

```
## [1] 0.1899117
```

```
# Rely on "Backward Selection" to obtain a model
ols_real_full <- lm(h1n1_vaccine ~., data = train)
step(ols_real_full, direction="backward")
```

```
## Start:  AIC=-18827.31
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_antiviral_meds +
##     behavioral_avoidance + behavioral_face_mask + behavioral_wash_hands +
##     behavioral_large_gatherings + behavioral_outside_home + behavioral_touch_face +
##     doctor_recc_h1n1 + chronic_med_condition + child_under_6_months +
##     health_worker + health_insurance + opinion_h1n1_vacc_effective +
##     opinion_h1n1_risk + opinion_h1n1_sick_from_vacc + age_group +
##     education + race + sex + income_poverty + employment_status +
##     household_adults + household_children
##
##                                 Df Sum of Sq    RSS    AIC
## - employment_status            3     0.215 1320.8 -18832
## - behavioral_outside_home      1     0.000 1320.6 -18829
## - household_adults             1     0.004 1320.6 -18829
## - household_children           1     0.004 1320.6 -18829
## - behavioral_antiviral_meds    1     0.012 1320.6 -18829
## - behavioral_wash_hands        1     0.015 1320.6 -18829
## - behavioral_face_mask         1     0.022 1320.6 -18829
## - h1n1_concern                 1     0.056 1320.6 -18829
## - behavioral_touch_face        1     0.064 1320.7 -18829
## - behavioral_avoidance         1     0.081 1320.7 -18829
## <none>                                     1320.6 -18827
## - behavioral_large_gatherings  1     0.402 1321.0 -18826
## - opinion_h1n1_sick_from_vacc  1     0.413 1321.0 -18826
## - chronic_med_condition        1     0.752 1321.3 -18824
## - income_poverty               3     1.662 1322.2 -18821
## - child_under_6_months         1     1.113 1321.7 -18821
## - education                    4     3.467 1324.0 -18810
## - race                         3     3.334 1323.9 -18809
## - h1n1_knowledge               1     3.110 1323.7 -18807
## - health_insurance             1     3.981 1324.6 -18801
## - sex                          1     4.632 1325.2 -18796
## - age_group                    4     8.125 1328.7 -18777
## - health_worker                1    18.718 1339.3 -18695
## - opinion_h1n1_vacc_effective  1    49.454 1370.0 -18478
## - opinion_h1n1_risk            1   108.004 1428.6 -18078
## - doctor_recc_h1n1             1   144.375 1465.0 -17838
##
## Step:  AIC=-18831.75
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_antiviral_meds +
##     behavioral_avoidance + behavioral_face_mask + behavioral_wash_hands +
##     behavioral_large_gatherings + behavioral_outside_home + behavioral_touch_face +
##     doctor_recc_h1n1 + chronic_med_condition + child_under_6_months +
##     health_worker + health_insurance + opinion_h1n1_vacc_effective +
##     opinion_h1n1_risk + opinion_h1n1_sick_from_vacc + age_group +
##     education + race + sex + income_poverty + household_adults +
##     household_children
##
##                               Df Sum of Sq    RSS    AIC
## - behavioral_outside_home      1     0.000 1320.8 -18834
## - household_adults             1     0.002 1320.8 -18834
## - household_children           1     0.004 1320.8 -18834
## - behavioral_antiviral_meds    1     0.010 1320.8 -18834
## - behavioral_wash_hands        1     0.016 1320.8 -18834
## - behavioral_face_mask         1     0.022 1320.8 -18834
## - h1n1_concern                 1     0.063 1320.9 -18833
## - behavioral_touch_face        1     0.065 1320.9 -18833
## - behavioral_avoidance         1     0.081 1320.9 -18833
## <none>                                   1320.8 -18832
## - behavioral_large_gatherings  1     0.398 1321.2 -18831
## - opinion_h1n1_sick_from_vacc  1     0.401 1321.2 -18831
## - chronic_med_condition        1     0.788 1321.6 -18828
## - child_under_6_months         1     1.102 1321.9 -18826
## - income_poverty               3     1.747 1322.5 -18825
## - education                    4     3.477 1324.3 -18815
## - race                         3     3.356 1324.2 -18814
## - h1n1_knowledge               1     3.098 1323.9 -18811
## - health_insurance             1     3.999 1324.8 -18805
## - sex                          1     4.615 1325.4 -18800
## - age_group                    4     9.875 1330.7 -18769
## - health_worker                1    18.924 1339.7 -18698
## - opinion_h1n1_vacc_effective  1    49.403 1370.2 -18483
## - opinion_h1n1_risk            1   108.058 1428.9 -18082
## - doctor_recc_h1n1             1   144.650 1465.5 -17841
##
## Step:  AIC=-18833.75
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_antiviral_meds +
##     behavioral_avoidance + behavioral_face_mask + behavioral_wash_hands +
##     behavioral_large_gatherings + behavioral_touch_face + doctor_recc_h1n1 +
##     chronic_med_condition + child_under_6_months + health_worker +
##     health_insurance + opinion_h1n1_vacc_effective + opinion_h1n1_risk +
##     opinion_h1n1_sick_from_vacc + age_group + education + race +
##     sex + income_poverty + household_adults + household_children
##
##                               Df Sum of Sq    RSS    AIC
## - household_adults            1     0.002 1320.8 -18836
## - household_children          1     0.004 1320.8 -18836
## - behavioral_antiviral_meds   1     0.010 1320.8 -18836
## - behavioral_wash_hands       1     0.016 1320.8 -18836
## - behavioral_face_mask        1     0.023 1320.8 -18836
## - h1n1_concern                1     0.063 1320.9 -18835
## - behavioral_touch_face       1     0.068 1320.9 -18835
## - behavioral_avoidance        1     0.080 1320.9 -18835
## <none>                                  1320.8 -18834
## - opinion_h1n1_sick_from_vacc  1     0.401 1321.2 -18833
## - behavioral_large_gatherings  1     0.504 1321.3 -18832
## - chronic_med_condition       1     0.789 1321.6 -18830
## - child_under_6_months        1     1.102 1321.9 -18828
## - income_poverty              3     1.747 1322.5 -18827
## - education                   4     3.487 1324.3 -18817
## - race                        3     3.359 1324.2 -18816
## - h1n1_knowledge              1     3.098 1323.9 -18813
## - health_insurance            1     3.998 1324.8 -18807
## - sex                         1     4.621 1325.4 -18802
## - age_group                   4     9.886 1330.7 -18770
## - health_worker               1    18.924 1339.7 -18700
## - opinion_h1n1_vacc_effective  1    49.408 1370.2 -18485
## - opinion_h1n1_risk           1   108.059 1428.9 -18084
## - doctor_recc_h1n1            1   144.649 1465.5 -17843
##
## Step:  AIC=-18835.73
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_antiviral_meds +
##     behavioral_avoidance + behavioral_face_mask + behavioral_wash_hands +
##     behavioral_large_gatherings + behavioral_touch_face + doctor_recc_h1n1 +
##     chronic_med_condition + child_under_6_months + health_worker +
##     health_insurance + opinion_h1n1_vacc_effective + opinion_h1n1_risk +
##     opinion_h1n1_sick_from_vacc + age_group + education + race +
##     sex + income_poverty + household_children
##
##                               Df Sum of Sq    RSS    AIC
## - household_children           1     0.004 1320.8 -18838
## - behavioral_antiviral_meds    1     0.010 1320.8 -18838
## - behavioral_wash_hands        1     0.016 1320.8 -18838
## - behavioral_face_mask         1     0.022 1320.8 -18838
## - h1n1_concern                 1     0.064 1320.9 -18837
## - behavioral_touch_face        1     0.067 1320.9 -18837
## - behavioral_avoidance         1     0.081 1320.9 -18837
## <none>                                   1320.8 -18836
## - opinion_h1n1_sick_from_vacc  1     0.400 1321.2 -18835
## - behavioral_large_gatherings  1     0.504 1321.3 -18834
## - chronic_med_condition        1     0.790 1321.6 -18832
## - child_under_6_months         1     1.100 1321.9 -18830
## - income_poverty               3     1.750 1322.5 -18829
## - education                    4     3.524 1324.3 -18818
## - race                         3     3.357 1324.2 -18818
## - h1n1_knowledge               1     3.096 1323.9 -18815
## - health_insurance             1     4.004 1324.8 -18809
## - sex                          1     4.621 1325.4 -18804
## - age_group                    4    10.178 1331.0 -18770
## - health_worker                1    18.927 1339.7 -18702
## - opinion_h1n1_vacc_effective  1    49.408 1370.2 -18487
## - opinion_h1n1_risk            1   108.062 1428.9 -18086
## - doctor_recc_h1n1             1   144.678 1465.5 -17845
##
## Step:  AIC=-18837.7
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_antiviral_meds +
##     behavioral_avoidance + behavioral_face_mask + behavioral_wash_hands +
##     behavioral_large_gatherings + behavioral_touch_face + doctor_recc_h1n1 +
##     chronic_med_condition + child_under_6_months + health_worker +
##     health_insurance + opinion_h1n1_vacc_effective + opinion_h1n1_risk +
##     opinion_h1n1_sick_from_vacc + age_group + education + race +
##     sex + income_poverty
##
##                               Df Sum of Sq    RSS    AIC
## - behavioral_antiviral_meds    1     0.010 1320.8 -18840
## - behavioral_wash_hands        1     0.016 1320.8 -18840
## - behavioral_face_mask         1     0.022 1320.8 -18840
## - h1n1_concern                 1     0.065 1320.9 -18839
## - behavioral_touch_face        1     0.068 1320.9 -18839
## - behavioral_avoidance         1     0.081 1320.9 -18839
## <none>                                   1320.8 -18838
## - opinion_h1n1_sick_from_vacc  1     0.401 1321.2 -18837
## - behavioral_large_gatherings  1     0.503 1321.3 -18836
## - chronic_med_condition        1     0.789 1321.6 -18834
## - child_under_6_months         1     1.114 1321.9 -18832
## - income_poverty               3     1.754 1322.6 -18831
## - education                    4     3.525 1324.3 -18820
## - race                         3     3.353 1324.2 -18820
## - h1n1_knowledge               1     3.109 1323.9 -18817
## - health_insurance             1     4.008 1324.8 -18811
## - sex                          1     4.621 1325.4 -18806
## - age_group                    4    11.715 1332.5 -18761
## - health_worker                1    18.924 1339.7 -18704
## - opinion_h1n1_vacc_effective  1    49.429 1370.2 -18489
## - opinion_h1n1_risk            1   108.115 1428.9 -18088
## - doctor_recc_h1n1             1   144.715 1465.5 -17846
##
## Step:  AIC=-18839.63
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_avoidance +
##     behavioral_face_mask + behavioral_wash_hands + behavioral_large_gatherings +
##     behavioral_touch_face + doctor_recc_h1n1 + chronic_med_condition +
##     child_under_6_months + health_worker + health_insurance +
##     opinion_h1n1_vacc_effective + opinion_h1n1_risk + opinion_h1n1_sick_from_vacc +
##     age_group + education + race + sex + income_poverty
##
##                               Df Sum of Sq    RSS    AIC
## - behavioral_wash_hands        1     0.016 1320.8 -18842
## - behavioral_face_mask         1     0.026 1320.8 -18841
## - h1n1_concern                 1     0.067 1320.9 -18841
## - behavioral_touch_face        1     0.068 1320.9 -18841
## - behavioral_avoidance         1     0.080 1320.9 -18841
## <none>                                   1320.8 -18840
## - opinion_h1n1_sick_from_vacc  1     0.400 1321.2 -18839
## - behavioral_large_gatherings  1     0.498 1321.3 -18838
## - chronic_med_condition        1     0.787 1321.6 -18836
## - child_under_6_months         1     1.115 1321.9 -18834
## - income_poverty               3     1.753 1322.6 -18833
## - education                    4     3.518 1324.3 -18822
## - race                         3     3.344 1324.2 -18822
## - h1n1_knowledge               1     3.113 1323.9 -18819
## - health_insurance             1     4.004 1324.8 -18813
## - sex                          1     4.625 1325.4 -18808
## - age_group                    4    11.720 1332.5 -18763
## - health_worker                1    18.921 1339.7 -18706
## - opinion_h1n1_vacc_effective  1    49.419 1370.2 -18491
## - opinion_h1n1_risk            1   108.539 1429.3 -18087
## - doctor_recc_h1n1             1   144.708 1465.5 -17848
##
## Step:  AIC=-18841.51
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_avoidance +
##     behavioral_face_mask + behavioral_large_gatherings + behavioral_touch_face +
##     doctor_recc_h1n1 + chronic_med_condition + child_under_6_months +
##     health_worker + health_insurance + opinion_h1n1_vacc_effective +
##     opinion_h1n1_risk + opinion_h1n1_sick_from_vacc + age_group +
##     education + race + sex + income_poverty
##
##                               Df Sum of Sq    RSS    AIC
## - behavioral_face_mask         1     0.026 1320.9 -18843
## - behavioral_touch_face        1     0.057 1320.9 -18843
## - h1n1_concern                 1     0.060 1320.9 -18843
## - behavioral_avoidance         1     0.100 1320.9 -18843
## <none>                                     1320.8 -18842
## - opinion_h1n1_sick_from_vacc  1     0.402 1321.2 -18841
## - behavioral_large_gatherings  1     0.506 1321.3 -18840
## - chronic_med_condition        1     0.789 1321.6 -18838
## - child_under_6_months         1     1.112 1321.9 -18836
## - income_poverty               3     1.756 1322.6 -18835
## - education                    4     3.534 1324.4 -18824
## - race                         3     3.374 1324.2 -18823
## - h1n1_knowledge               1     3.101 1323.9 -18821
## - health_insurance             1     3.993 1324.8 -18815
## - sex                          1     4.697 1325.5 -18810
## - age_group                    4    11.733 1332.6 -18765
## - health_worker                1    18.907 1339.7 -18708
## - opinion_h1n1_vacc_effective  1    49.425 1370.3 -18493
## - opinion_h1n1_risk            1   108.527 1429.4 -18089
## - doctor_recc_h1n1             1   144.704 1465.5 -17850
##
## Step:  AIC=-18843.33
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_avoidance +
##     behavioral_large_gatherings + behavioral_touch_face + doctor_recc_h1n1 +
##     chronic_med_condition + child_under_6_months + health_worker +
##     health_insurance + opinion_h1n1_vacc_effective + opinion_h1n1_risk +
##     opinion_h1n1_sick_from_vacc + age_group + education + race +
##     sex + income_poverty
##
##                               Df Sum of Sq    RSS    AIC
## - behavioral_touch_face        1     0.060 1320.9 -18845
## - h1n1_concern                 1     0.066 1320.9 -18845
## - behavioral_avoidance         1     0.100 1321.0 -18845
## <none>                                     1320.9 -18843
## - opinion_h1n1_sick_from_vacc  1     0.398 1321.2 -18842
## - behavioral_large_gatherings  1     0.486 1321.3 -18842
## - chronic_med_condition        1     0.799 1321.7 -18840
## - child_under_6_months         1     1.118 1322.0 -18837
## - income_poverty               3     1.752 1322.6 -18837
## - education                    4     3.521 1324.4 -18826
## - race                         3     3.371 1324.2 -18825
## - h1n1_knowledge               1     3.131 1324.0 -18823
## - health_insurance             1     3.978 1324.8 -18817
## - sex                          1     4.692 1325.5 -18812
## - age_group                    4    11.738 1332.6 -18767
## - health_worker                1    19.015 1339.9 -18709
## - opinion_h1n1_vacc_effective  1    49.400 1370.3 -18495
## - opinion_h1n1_risk            1   108.929 1429.8 -18088
## - doctor_recc_h1n1             1   144.813 1465.7 -17852
##
## Step:  AIC=-18844.9
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_avoidance +
##     behavioral_large_gatherings + doctor_recc_h1n1 + chronic_med_condition +
##     child_under_6_months + health_worker + health_insurance +
##     opinion_h1n1_vacc_effective + opinion_h1n1_risk + opinion_h1n1_sick_from_vacc +
##     age_group + education + race + sex + income_poverty
##
##                               Df Sum of Sq    RSS    AIC
## - behavioral_avoidance         1     0.068 1321.0 -18846
## - h1n1_concern                 1     0.078 1321.0 -18846
## <none>                                     1320.9 -18845
## - opinion_h1n1_sick_from_vacc  1     0.393 1321.3 -18844
## - behavioral_large_gatherings  1     0.449 1321.4 -18844
## - chronic_med_condition        1     0.788 1321.7 -18841
## - child_under_6_months         1     1.123 1322.0 -18839
## - income_poverty               3     1.751 1322.7 -18838
## - education                    4     3.494 1324.4 -18828
## - race                         3     3.344 1324.3 -18827
## - h1n1_knowledge               1     3.174 1324.1 -18824
## - health_insurance             1     3.983 1324.9 -18818
## - sex                          1     4.632 1325.5 -18814
## - age_group                    4    11.782 1332.7 -18768
## - health_worker                1    19.146 1340.1 -18709
## - opinion_h1n1_vacc_effective  1    49.452 1370.4 -18496
## - opinion_h1n1_risk            1   109.272 1430.2 -18088
## - doctor_recc_h1n1             1   144.981 1465.9 -17852
##
## Step:  AIC=-18846.4
## h1n1_vaccine ~ h1n1_concern + h1n1_knowledge + behavioral_large_gatherings +
##     doctor_recc_h1n1 + chronic_med_condition + child_under_6_months +
##     health_worker + health_insurance + opinion_h1n1_vacc_effective +
##     opinion_h1n1_risk + opinion_h1n1_sick_from_vacc + age_group +
##     education + race + sex + income_poverty
##
```

```
##                               Df Sum of Sq    RSS    AIC
## - h1n1_concern                1     0.059 1321.0 -18848
## <none>                                     1321.0 -18846
## - opinion_h1n1_sick_from_vacc  1     0.408 1321.4 -18846
## - behavioral_large_gatherings  1     0.537 1321.5 -18844
## - chronic_med_condition        1     0.789 1321.8 -18843
## - child_under_6_months         1     1.132 1322.1 -18840
## - income_poverty               3     1.756 1322.7 -18840
## - education                    4     3.475 1324.5 -18829
## - race                         3     3.340 1324.3 -18828
## - h1n1_knowledge               1     3.144 1324.1 -18826
## - health_insurance             1     3.959 1324.9 -18820
## - sex                          1     4.801 1325.8 -18814
## - age_group                    4    11.876 1332.9 -18769
## - health_worker                1    19.236 1340.2 -18710
## - opinion_h1n1_vacc_effective  1    49.384 1370.4 -18498
## - opinion_h1n1_risk            1   109.276 1430.3 -18089
## - doctor_recc_h1n1             1   144.916 1465.9 -17854
## 
## Step:  AIC=-18847.98
## h1n1_vaccine ~ h1n1_knowledge + behavioral_large_gatherings + 
##     doctor_recc_h1n1 + chronic_med_condition + child_under_6_months + 
##     health_worker + health_insurance + opinion_h1n1_vacc_effective + 
##     opinion_h1n1_risk + opinion_h1n1_sick_from_vacc + age_group + 
##     education + race + sex + income_poverty
## 
##                               Df Sum of Sq    RSS    AIC
## <none>                                     1321.0 -18848
## - opinion_h1n1_sick_from_vacc  1     0.360 1321.4 -18847
## - behavioral_large_gatherings  1     0.495 1321.5 -18846
## - chronic_med_condition        1     0.795 1321.8 -18844
## - child_under_6_months         1     1.149 1322.2 -18842
## - income_poverty               3     1.758 1322.8 -18841
## - education                    4     3.454 1324.5 -18831
## - race                         3     3.283 1324.3 -18830
## - h1n1_knowledge               1     3.193 1324.2 -18827
## - health_insurance             1     3.949 1325.0 -18822
## - sex                          1     4.744 1325.8 -18816
## - age_group                    4    12.008 1333.0 -18770
## - health_worker                1    19.196 1340.2 -18712
## - opinion_h1n1_vacc_effective  1    52.071 1373.1 -18481
## - opinion_h1n1_risk            1   114.888 1435.9 -18053
## - doctor_recc_h1n1             1   145.193 1466.2 -17854

## 
## Call:
## lm(formula = h1n1_vaccine ~ h1n1_knowledge + behavioral_large_gatherings + 
##     doctor_recc_h1n1 + chronic_med_condition + child_under_6_months + 
##     health_worker + health_insurance + opinion_h1n1_vacc_effective + 
##     opinion_h1n1_risk + opinion_h1n1_sick_from_vacc + age_group + 
##     education + race + sex + income_poverty, data = train)
## 
## Coefficients:
##                              (Intercept)
##                               -0.448728
##                          h1n1_knowledge
##                                0.033018
##             behavioral_large_gatherings
##                               -0.016019
##                        doctor_recc_h1n1
##                                0.304049
##                   chronic_med_condition
##                                0.021077
##                    child_under_6_months
##                                0.040238
##                           health_worker
##                                0.145893
##                        health_insurance
##                                0.067628
##             opinion_h1n1_vacc_effective
##                                0.078603
##                       opinion_h1n1_risk
##                                0.098885
##             opinion_h1n1_sick_from_vacc
##                               -0.004875
##                  age_group35 - 44 Years
##                               -0.002282
##                  age_group45 - 54 Years
##                               -0.004553
##                  age_group55 - 64 Years
##                                0.060859
##                      age_group65+ Years
##                                0.083684
##                     education< 12 Years
##                               -0.108644
##                       education12 Years
##                               -0.075570
##               educationCollege Graduate
##                               -0.038556
##                   educationSome College
##                               -0.063330
##                            raceHispanic
##                                0.058458
##                   raceOther or Multiple
##                                0.077748
##                               raceWhite
##                                0.069645
##                                 sexMale
##                                0.046730
## income_poverty<= $75,000, Above Poverty
##                               -0.039933
##                 income_poverty> $75,000
##                               -0.024643
##             income_povertyBelow Poverty
##                               -0.049307
```

** PART II – LOGISTIC REGRESSION ****************************** Logistic regression is used for modeling categorical outcomes, particularly no/yes, 0/1 outcomes.Such problems are called (binary) classification problems. Under the consideration of dependent variable “h1n1_vaccine”, we decide to use logistic regression with binary classification.

```r
# Apply glm() -- logistic regression model
glm <- glm(h1n1_vaccine ~.,family = "binomial", data = train)
summary(glm)
```

```
##
## Call:
## glm(formula = h1n1_vaccine ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6118  -0.6292  -0.3853   0.5196   3.4694
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -6.603912   0.497876 -13.264  < 2e-16
## h1n1_concern                             0.047383   0.038365   1.235 0.216812
## h1n1_knowledge                           0.223705   0.051164   4.372 1.23e-05
## behavioral_antiviral_meds                0.054936   0.128770   0.427 0.669654
## behavioral_avoidance                    -0.052921   0.072266  -0.732 0.463977
## behavioral_face_mask                     0.055617   0.112189   0.496 0.620076
## behavioral_wash_hands                    0.008729   0.086464   0.101 0.919589
## behavioral_large_gatherings             -0.132883   0.074359  -1.787 0.073930
## behavioral_outside_home                 -0.011669   0.075668  -0.154 0.877445
## behavioral_touch_face                    0.034178   0.069274   0.493 0.621754
## doctor_recc_h1n1                         1.638919   0.061523  26.639  < 2e-16
## chronic_med_condition                    0.131787   0.063404   2.079 0.037661
## child_under_6_months                     0.261787   0.101389   2.582 0.009823
## health_worker                            0.896370   0.086426  10.372  < 2e-16
## health_insurance                         0.679779   0.109768   6.193 5.91e-10
## opinion_h1n1_vacc_effective              0.715664   0.037826  18.920  < 2e-16
## opinion_h1n1_risk                        0.571238   0.024547  23.271  < 2e-16
## opinion_h1n1_sick_from_vacc             -0.021035   0.023164  -0.908 0.363831
## age_group35 - 44 Years                  -0.067294   0.105293  -0.639 0.522748
## age_group45 - 54 Years                  -0.093011   0.097244  -0.956 0.338831
## age_group55 - 64 Years                   0.373615   0.096551   3.870 0.000109
## age_group65+ Years                       0.509887   0.104186   4.894 9.88e-07
## education< 12 Years                     -0.914986   0.350470  -2.611 0.009035
## education12 Years                       -0.636020   0.339101  -1.876 0.060710
## educationCollege Graduate               -0.386693   0.337919  -1.144 0.252483
## educationSome College                   -0.557993   0.338334  -1.649 0.099099
## raceHispanic                             0.565242   0.163073   3.466 0.000528
## raceOther or Multiple                    0.706712   0.161840   4.367 1.26e-05
## raceWhite                                0.616345   0.121302   5.081 3.75e-07
## sexMale                                  0.346906   0.060714   5.714 1.10e-08
## income_poverty<= $75,000, Above Poverty -0.303538   0.096446  -3.147 0.001648
## income_poverty> $75,000                 -0.196620   0.108964  -1.804 0.071162
## income_povertyBelow Poverty             -0.423226   0.132401  -3.197 0.001391
## employment_statusEmployed               -0.320311   0.345961  -0.926 0.354520
## employment_statusNot in Labor Force     -0.265648   0.346048  -0.768 0.442688
## employment_statusUnemployed             -0.313165   0.365357  -0.857 0.391364
## household_adults                        -0.014680   0.039317  -0.373 0.708873
## household_children                       0.011766   0.037251   0.316 0.752106
##
## (Intercept)                             ***
## h1n1_concern
## h1n1_knowledge                          ***
## behavioral_antiviral_meds
## behavioral_avoidance
## behavioral_face_mask
## behavioral_wash_hands
## behavioral_large_gatherings             .
## behavioral_outside_home
## behavioral_touch_face
## doctor_recc_h1n1                        ***
## chronic_med_condition                   *
## child_under_6_months                    **
## health_worker                           ***
## health_insurance                        ***
## opinion_h1n1_vacc_effective             ***
## opinion_h1n1_risk                       ***
## opinion_h1n1_sick_from_vacc
## age_group35 - 44 Years
## age_group45 - 54 Years
## age_group55 - 64 Years                  ***
## age_group65+ Years                      ***
## education< 12 Years                     **
## education12 Years                       .
## educationCollege Graduate
## educationSome College                   .
## raceHispanic                            ***
## raceOther or Multiple                   ***
## raceWhite                               ***
## sexMale                                 ***
## income_poverty<= $75,000, Above Poverty **
## income_poverty> $75,000                 .
## income_povertyBelow Poverty             **
## employment_statusEmployed
## employment_statusNot in Labor Force
## employment_statusUnemployed
## household_adults
## household_children
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11587.1  on 9552  degrees of freedom
## Residual deviance:  8023.4  on 9515  degrees of freedom
## AIC: 8099.4
##
## Number of Fisher Scoring iterations: 5
```

```
yhat_glm <- predict(glm, type="response")

# confusion matrix
```

```
table(train$h1n1_vaccine, (yhat_glm>0.5))
```

```
##
##     FALSE TRUE
##   0  6191  545
##   1  1180 1637
```

```
TPR(train$h1n1_vaccine,(yhat_glm>0.5))
```

```
## [1] 0.5811147
```

```
TNR(train$h1n1_vaccine,(yhat_glm>0.5))
```

```
## [1] 0.9190914
```

```
ME(train$h1n1_vaccine,(yhat_glm>0.5))
```

```
## [1] 0.1805715
```

```
yhat.glm <- predict(glm, test, type="response")
table(test$h1n1_vaccine, (yhat.glm > 0.5))
```

```
##
##     FALSE TRUE
##   0  2528  218
##   1   530  689
```

```
TPR(test$h1n1_vaccine, (yhat.glm > 0.5))
```

```
## [1] 0.5652174
```

```
TNR(test$h1n1_vaccine, (yhat.glm > 0.5))
```

```
## [1] 0.9206118
```

```
# Misclassification Error
ME(test$h1n1_vaccine, (yhat.glm > 0.5))
```

```
## [1] 0.1886507
```

```
# LDA/QDA
lda <- lda(h1n1_vaccine ~ ., data=train)
yhat_lda <- predict(lda)$posterior[,2]
table(train$h1n1_vaccine, (yhat_lda >0.5))
```

```
##
##     FALSE TRUE
##   0  6143  593
##   1  1173 1644
```

```
yhat.lda <- predict(lda, test)$posterior[,2]
table(test$h1n1_vaccine, (yhat.lda > 0.5))
```

```
##
##     FALSE TRUE
##   0  2519  227
##   1   523  696
```

```
TPR(test$h1n1_vaccine, (yhat.lda >0.5))
```

```
## [1] 0.5709598
```

```r
TNR(test$h1n1_vaccine, (yhat.lda >0.5))
```

```
## [1] 0.9173343
```

```r
ME(test$h1n1_vaccine, (yhat.lda >0.5))
```

```
## [1] 0.1891551
```

```r
qda <- qda(h1n1_vaccine ~ ., data=train)
yhat_qda <- predict(qda)$posterior[,2]
table(train$h1n1_vaccine, (yhat_qda >0.5))
```

```
##
##     FALSE TRUE
##   0  5508 1228
##   1   912 1905
```

```r
yhat.qda <- predict(qda, test)$posterior[,2]
table(test$h1n1_vaccine, (yhat.qda > 0.5))
```

```
##
##     FALSE TRUE
##   0  2269  477
##   1   401  818
```

```r
TPR(test$h1n1_vaccine, (yhat.qda >0.5))
```

```
## [1] 0.6710418
```

```r
TNR(test$h1n1_vaccine, (yhat.qda >0.5))
```

```
## [1] 0.8262928
```

```r
ME(test$h1n1_vaccine, (yhat.qda >0.5))
```

```
## [1] 0.2214376
```

```r
# Plot ROC graph
par(mfrow=c(1,1))
glm.roc <- roc(test$h1n1_vaccine, yhat.glm, direction = "<")
```

```
## Setting levels: control = 0, case = 1
```

```r
glm.roc
```

```
##
## Call:
## roc.default(response = test$h1n1_vaccine, predictor = yhat.glm,     direction = "<")
##
## Data: yhat.glm in 2746 controls (test$h1n1_vaccine 0) < 1219 cases (test$h1n1_vaccine 1).
## Area under the curve: 0.8535
```

```r
lda.roc <- roc(test$h1n1_vaccine, yhat.lda, direction = "<")
```

```
## Setting levels: control = 0, case = 1
```

```r
lda.roc
```

```
##
## Call:
## roc.default(response = test$h1n1_vaccine, predictor = yhat.lda,     direction = "<")
```

```
##
## Data: yhat.lda in 2746 controls (test$h1n1_vaccine 0) < 1219 cases (test$h1n1_vaccine 1).
## Area under the curve: 0.8534
```

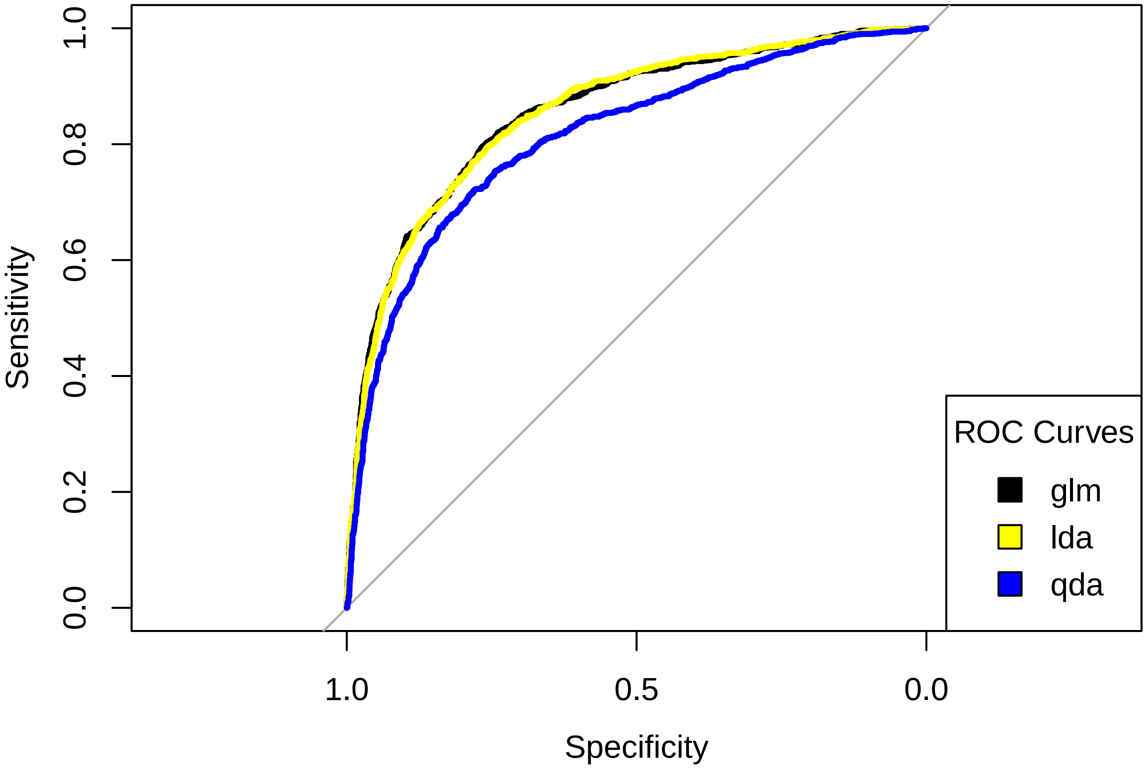
```
qda.roc <- roc(test$h1n1_vaccine, yhat.qda, direction = "<")
```

```
## Setting levels: control = 0, case = 1
```

```
qda.roc
```

```
##
## Call:
## roc.default(response = test$h1n1_vaccine, predictor = yhat.qda,     direction = "<")
##
## Data: yhat.qda in 2746 controls (test$h1n1_vaccine 0) < 1219 cases (test$h1n1_vaccine 1).
## Area under the curve: 0.8127
```

```
plot(glm.roc, lwd=3)
lines(lda.roc, lwd=3, col = "yellow")
lines(qda.roc, lwd=3, col = "blue")
legend("bottomright",title="ROC Curves",c("glm","lda","qda"), fill=c("black","yellow","blue"))
```



** PART III – Classification Tree ******************************

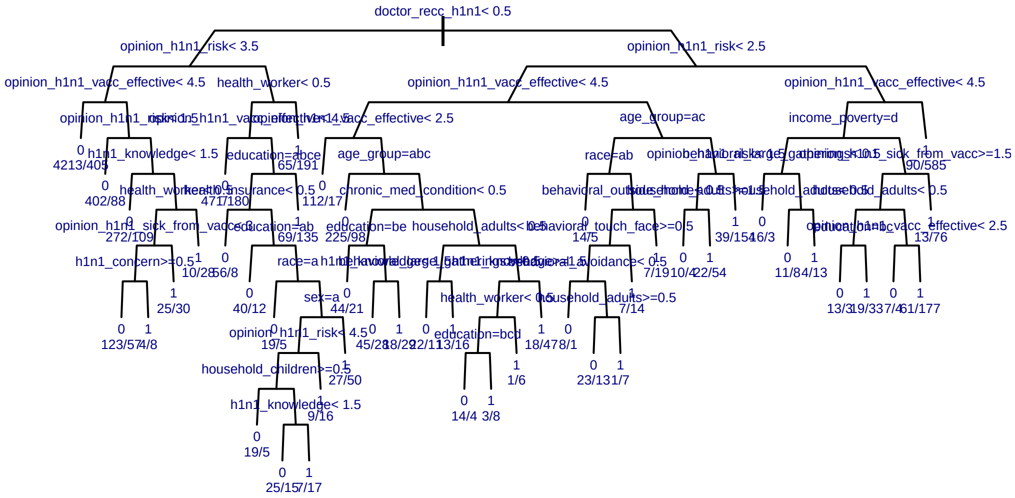
```
paste(colnames(train[,-c(1)]), collapse = "+")
```

```
## [1] "h1n1_knowledge+behavioral_antiviral_meds+behavioral_avoidance+behavioral_face_mask+behavioral_wa
```

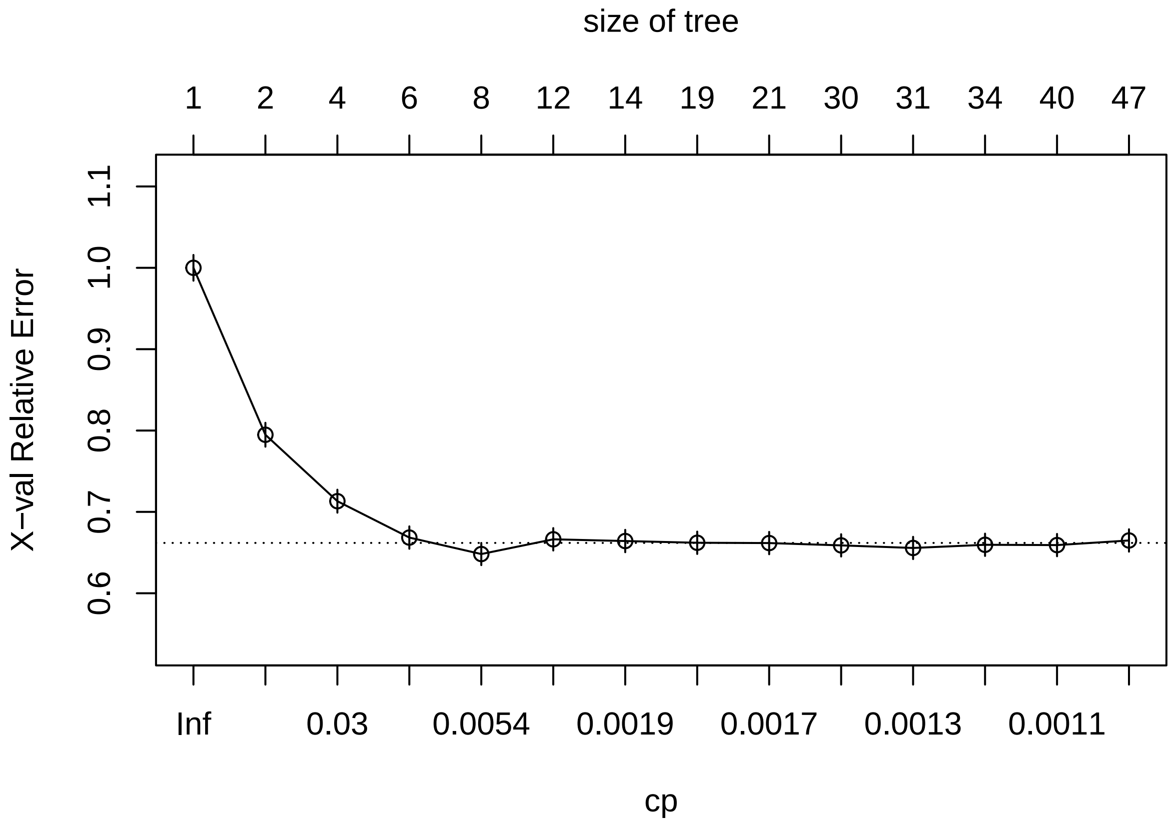
```
form1 <- formula(h1n1_vaccine ~ h1n1_concern+h1n1_knowledge+behavioral_antiviral_meds+behavioral_avoida

t1 <- rpart(form1, data=train, cp=.001, method="class")
```

```
plot(t1,uniform=T,compress=T,margin=.05,branch=0.9)
text(t1, cex=.4, col="navy",use.n=TRUE)
```



```
plotcp(t1) #plot cross-validation results
```



```
CP <- printcp(t1) #display cp table
```

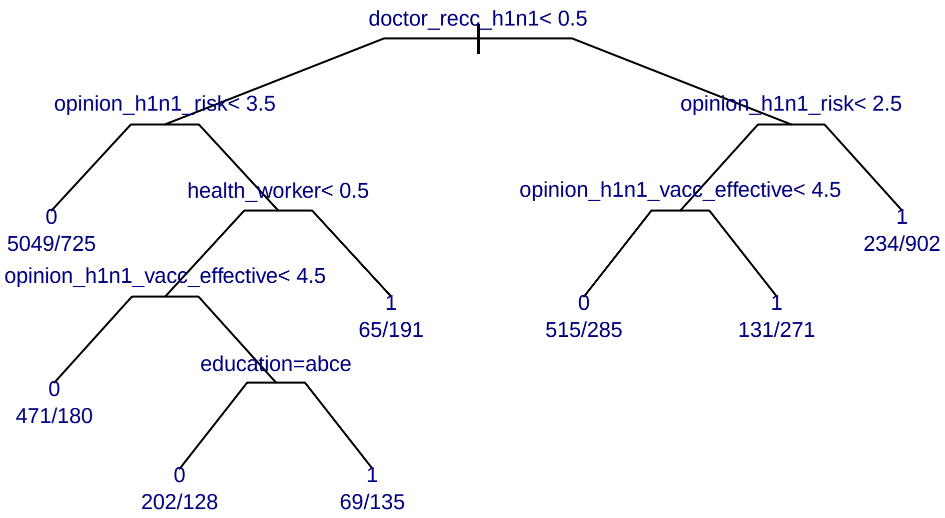
```
##
## Classification tree:
## rpart(formula = form1, data = train, method = "class", cp = 0.001)
##
## Variables actually used in tree construction:
```

```
##  [1] age_group                   behavioral_avoidance
##  [3] behavioral_large_gatherings behavioral_outside_home
##  [5] behavioral_touch_face       chronic_med_condition
##  [7] doctor_recc_h1n1            education
##  [9] h1n1_concern                h1n1_knowledge
## [11] health_insurance            health_worker
## [13] household_adults            household_children
## [15] income_poverty              opinion_h1n1_risk
## [17] opinion_h1n1_sick_from_vacc opinion_h1n1_vacc_effective
## [19] race                        sex
##
## Root node error: 2817/9553 = 0.29488
##
## n= 9553
##
##           CP nsplit rel error  xerror     xstd
## 1  0.2051828      0   1.00000 1.00000 0.015821
## 2  0.0408236      1   0.79482 0.79482 0.014698
## 3  0.0223642      3   0.71317 0.71317 0.014140
## 4  0.0117146      5   0.66844 0.66844 0.013803
## 5  0.0024849      7   0.64501 0.64821 0.013643
## 6  0.0019524     11   0.63472 0.66631 0.013786
## 7  0.0018933     13   0.63081 0.66418 0.013769
## 8  0.0017749     18   0.62016 0.66205 0.013753
## 9  0.0015974     20   0.61661 0.66170 0.013750
## 10 0.0014200     29   0.60135 0.65886 0.013728
## 11 0.0012425     30   0.59993 0.65566 0.013702
## 12 0.0011833     33   0.59531 0.65957 0.013733
## 13 0.0010650     39   0.58821 0.65921 0.013730
## 14 0.0010000     46   0.58076 0.66489 0.013775
```

```r
cp <- CP[,1][CP[,4] == min(CP[,4])]
cp
```

```
##           5
## 0.002484913
```

```r
t2 <- prune(t1,cp=cp[1])
plot(t2,uniform=T,compress=T,margin=.05,branch=0.3)
text(t2, cex=.7, col="navy",use.n=TRUE)
```



```
# Predictions
yhat.t2 <- predict(t2, test, type="prob")[,2]
table(test$h1n1_vaccine, (yhat.t2>0.5))
```

```
##
##     FALSE TRUE
##   0  2556  190
##   1   590  629
```

```
TPR(test$h1n1_vaccine,(yhat.t2>0.5))
```

```
## [1] 0.5159967
```

```
TNR(test$h1n1_vaccine,(yhat.t2>0.5))
```

```
## [1] 0.9308084
```

```
ME(test$h1n1_vaccine,(yhat.t2>0.5))
```

```
## [1] 0.1967213
```

```
tree.roc <- roc(test$h1n1_vaccine, yhat.t2 , direction="<")
```

```
## Setting levels: control = 0, case = 1
```

```
tree.roc
```

```
##
## Call:
## roc.default(response = test$h1n1_vaccine, predictor = yhat.t2,     direction = "<")
##
## Data: yhat.t2 in 2746 controls (test$h1n1_vaccine 0) < 1219 cases (test$h1n1_vaccine 1).
## Area under the curve: 0.7979
```

** PART IV – Random Forest ******************************

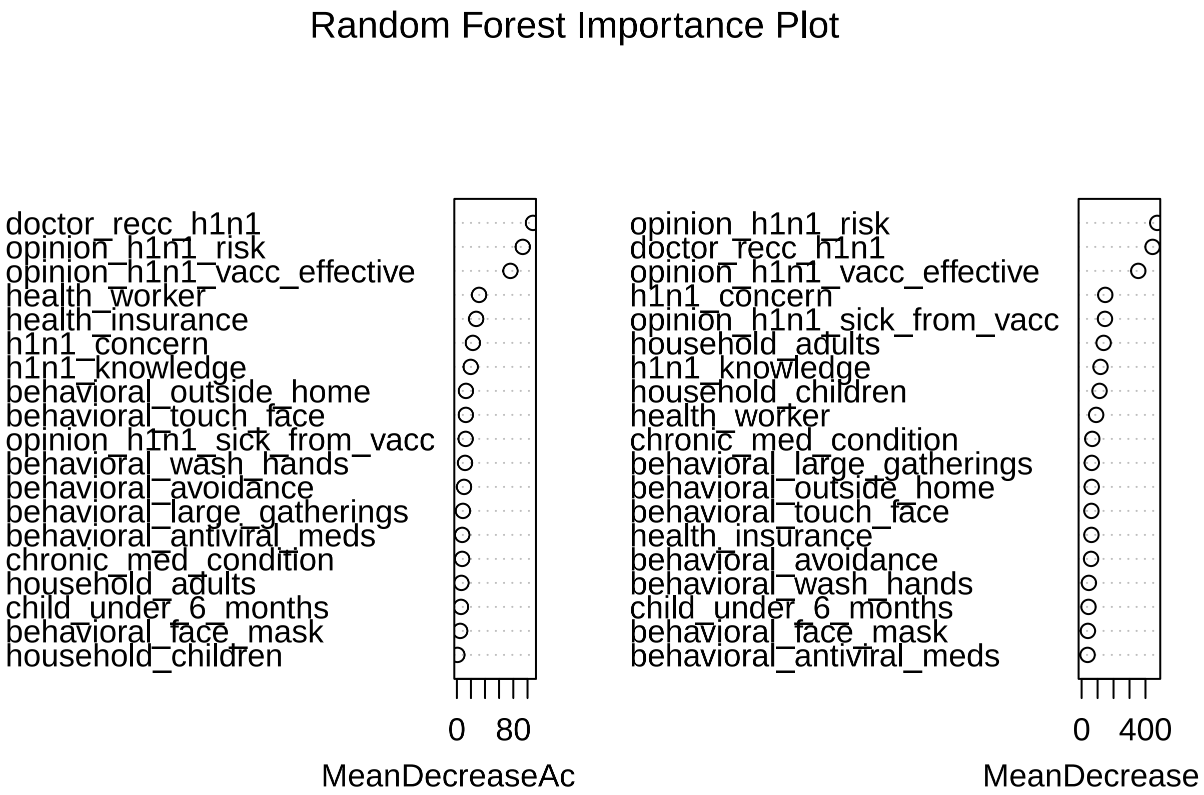
```
train_rf <- train[,-c(18:23)]
X <- as.matrix(train_rf[,-c(20)])
Y <- factor(train_rf$h1n1_vaccine)
rf1 <- randomForest(x=X, y=Y, data=train_rf, ntree=500, mtry=3, importance=T, na.action=na.omit)
summary(rf1)
```

```
##                 Length Class  Mode     
## call                8  -none- call     
## type                1  -none- character
## predicted        9553  factor numeric  
## err.rate         1500  -none- numeric  
## confusion           6  -none- numeric  
## votes           19106  matrix numeric  
## oob.times        9553  -none- numeric  
## classes             2  -none- character
## importance         76  -none- numeric  
## importanceSD       57  -none- numeric  
## localImportance     0  -none- NULL     
## proximity           0  -none- NULL     
## ntree               1  -none- numeric  
## mtry                1  -none- numeric  
## forest             14  -none- list     
## y                9553  factor numeric  
## test                0  -none- NULL     
## inbag               0  -none- NULL     
```

```
head(rf1$importance)
```

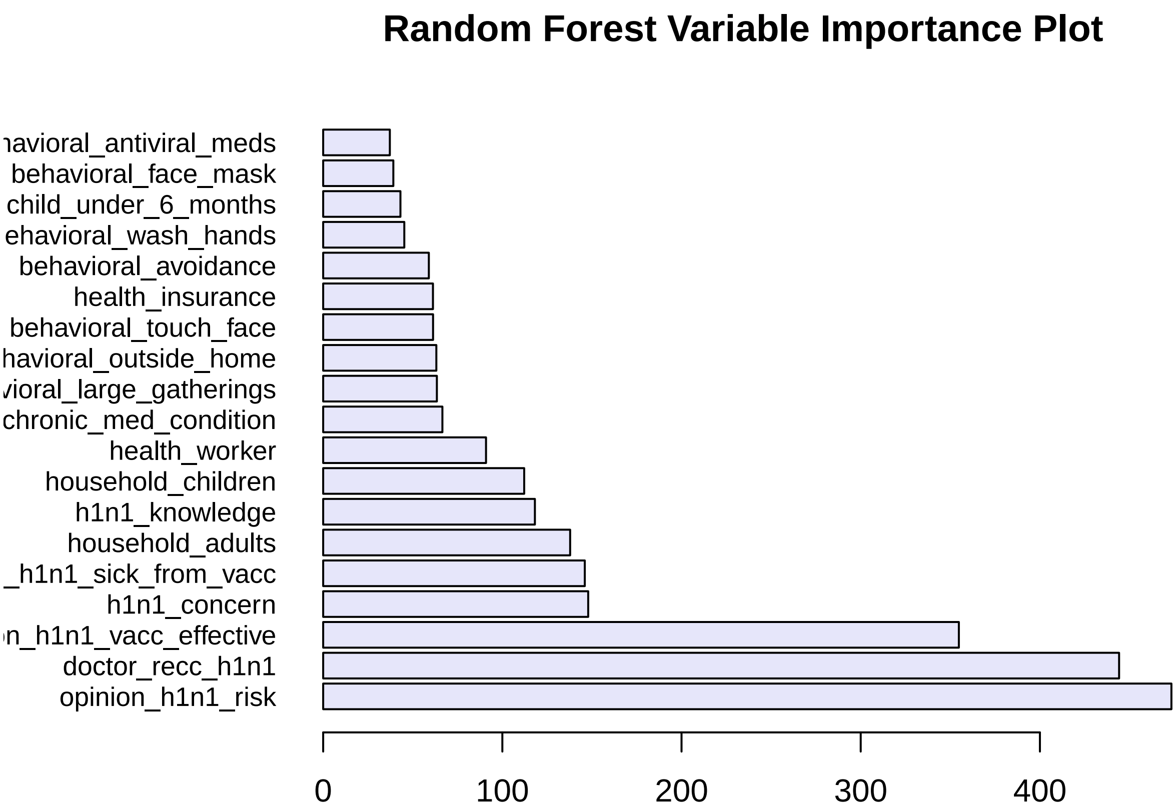
```
##                                      0             1 MeanDecreaseAccuracy
## h1n1_concern              0.0090740264 -2.327036e-03         0.0057093587
## h1n1_knowledge            0.0013622223  1.215420e-02         0.0045439015
## behavioral_antiviral_meds 0.0009183709 -3.282975e-05         0.0006370152
## behavioral_avoidance      0.0014792118  2.146473e-03         0.0016757294
## behavioral_face_mask      0.0011508971 -1.175801e-03         0.0004627176
## behavioral_wash_hands     0.0024378189  7.434185e-05         0.0017431574
##                           MeanDecreaseGini
## h1n1_concern                     147.91869
## h1n1_knowledge                   118.16785
## behavioral_antiviral_meds         37.20778
## behavioral_avoidance              58.97319
## behavioral_face_mask              39.15975
## behavioral_wash_hands             45.27524
```

```
# variable importance ranking 1
varImpPlot(rf1, sort = TRUE , main = "Random Forest Importance Plot")
```



```
# variable importance ranking 2
imp <- rf1$importance[,4]
ord <- order(imp, decreasing=T)
imp <- imp[ord]

par(mar=c(2, 8, 4, 2) + 0.1)
barplot(imp, col='lavender', horiz=TRUE, las=1, cex.names=.8)
title("Random Forest Variable Importance Plot")
```



```
# Predictions
yhat.rf <- predict(rf1, test, type="prob")[,2]
table(test$h1n1_vaccine, (yhat.rf>0.5))
```

```
##
##     FALSE TRUE
##   0  2532  214
##   1   551  668
```

```
TPR(test$h1n1_vaccine,(yhat.rf>0.5))
```

```
## [1] 0.5479902
```

```
TNR(test$h1n1_vaccine,(yhat.rf>0.5))
```

```
## [1] 0.9220685
```

```
ME(test$h1n1_vaccine,(yhat.rf>0.5))
```

```
## [1] 0.1929382
```

```
# Plot ROC graph
par(mfrow=c(1,1))
rf.roc <- roc(test$h1n1_vaccine, yhat.rf, direction="<")
```

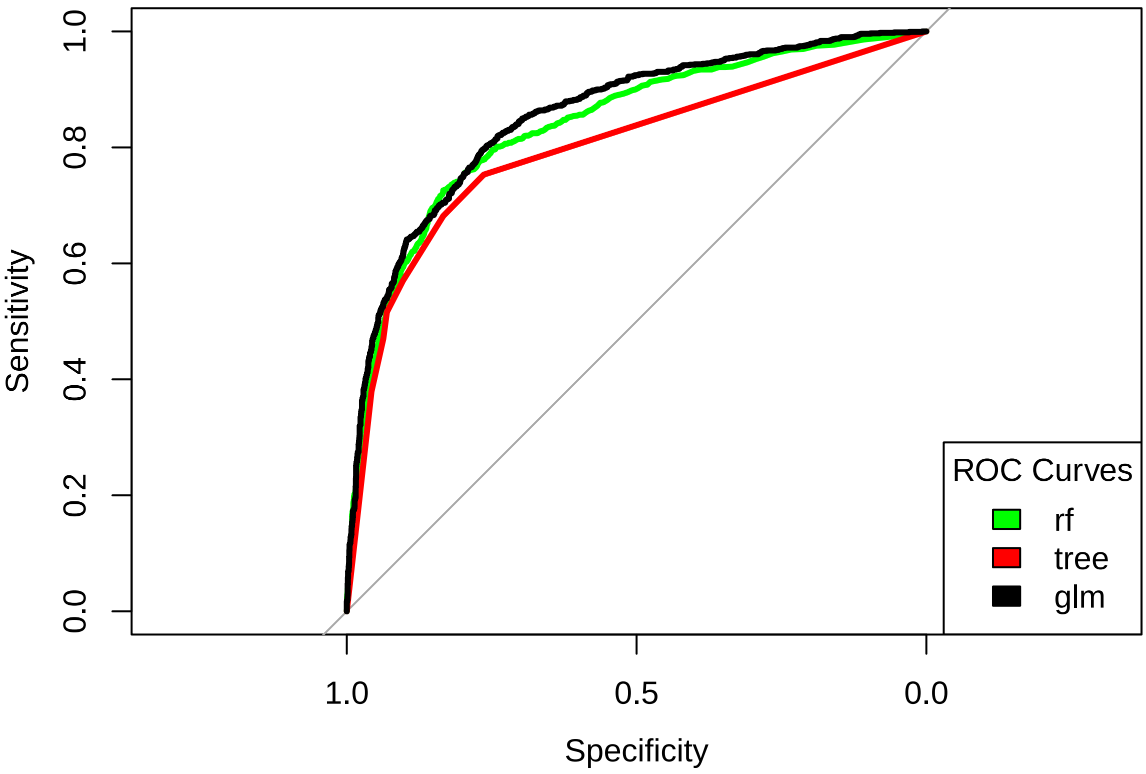
```
## Setting levels: control = 0, case = 1
```

```
rf.roc
```

```
##
## Call:
## roc.default(response = test$h1n1_vaccine, predictor = yhat.rf,     direction = "<")
##
```

```
## Data: yhat.rf in 2746 controls (test$h1n1_vaccine 0) < 1219 cases (test$h1n1_vaccine 1).
## Area under the curve: 0.8386
```

```
plot(rf.roc, lwd=3, col = "green")
lines(tree.roc, lwd=3, col = "red")
lines(glm.roc, lwd=3)
legend("bottomright",title="ROC Curves",c("rf","tree","glm"), fill=c("green","red","black"))
```



** PART V – Neural Nets ********************************

```
data_nn <- data[,-c(1,19:24)]
data_nn[,-c(20)] <- scale(data_nn[,-c(20)])
data_nn$h1n1_vaccine <- factor(data_nn$h1n1_vaccine)
train2 <- data_nn[trn==TRUE,]
test2 <- data_nn[trn==FALSE,]

paste(colnames(data_nn), collapse = "+")
```

```
## [1] "h1n1_concern+h1n1_knowledge+behavioral_antiviral_meds+behavioral_avoidance+behavioral_face_mask
```

```
form2 <- h1n1_vaccine ~ h1n1_concern+h1n1_knowledge+behavioral_antiviral_meds+behavioral_avoidance+beha

n1 <- nnet(form2, data = train2, size = 7, maxit = 500, decay=0.002)
```

```
## # weights:  148
## initial  value 7178.674103
## iter  10 value 4393.853511
## iter  20 value 4087.262121
## iter  30 value 3994.128123
## iter  40 value 3967.558016
## iter  50 value 3959.333187
## iter  60 value 3952.044998
## iter  70 value 3945.539572
## iter  80 value 3938.587564
## iter  90 value 3933.574834
## iter 100 value 3927.759861
## iter 110 value 3924.918517
## iter 120 value 3924.031797
## iter 130 value 3923.708109
## iter 140 value 3922.652410
## iter 150 value 3921.838955
## iter 160 value 3921.335043
## iter 170 value 3921.086963
## iter 180 value 3920.453649
## iter 190 value 3920.279914
## iter 200 value 3919.734625
## iter 210 value 3919.001932
## iter 220 value 3917.940512
## iter 230 value 3917.170019
## iter 240 value 3916.662917
## iter 250 value 3916.323292
## iter 260 value 3916.220907
## iter 270 value 3916.035838
## iter 280 value 3916.025299
## final  value 3916.022232
## converged
```

```
# Predictions
yhat.n1 <- predict(n1, test)
table(test$h1n1_vaccine, (yhat.n1>0.5))
```

```
##
##     FALSE TRUE
##   0     9 2737
##   1     0 1219
```

```
TPR(test$h1n1_vaccine,(yhat.n1>0.5))
```

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

```
TNR(test$h1n1_vaccine,(yhat.n1>0.5))
```

```
## [1] 0.003277495
```

```
ME(test$h1n1_vaccine,(yhat.n1>0.5))
```

```
## [1] 0.69029
```

```
# Plot ROC graph
par(mfrow=c(1,1))
nn.roc <- roc(test$h1n1_vaccine, yhat.n1, direction="<")
```

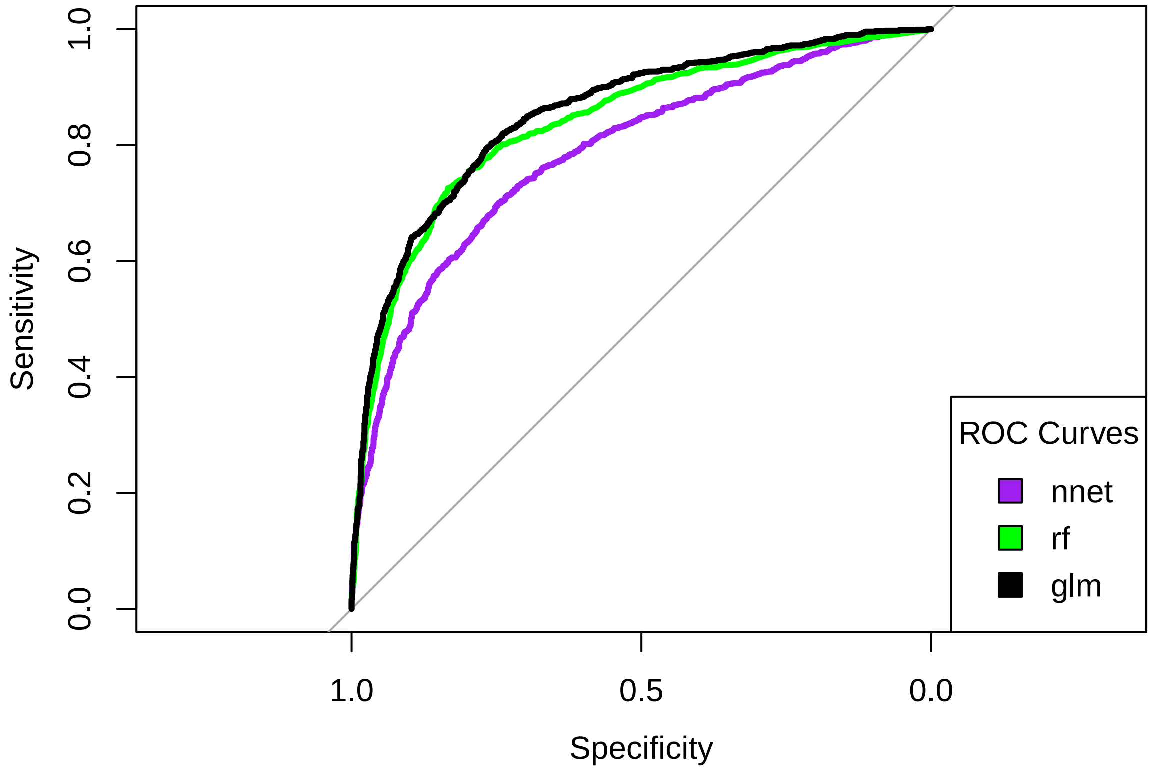
```
## Setting levels: control = 0, case = 1
```

```
## Warning in roc.default(test$h1n1_vaccine, yhat.n1, direction = "<"): Deprecated
## use a matrix as predictor. Unexpected results may be produced, please pass a
## numeric vector.
```

```
nn.roc
```

```
##
## Call:
## roc.default(response = test$h1n1_vaccine, predictor = yhat.n1,     direction = "<")
##
## Data: yhat.n1 in 2746 controls (test$h1n1_vaccine 0) < 1219 cases (test$h1n1_vaccine 1).
## Area under the curve: 0.7855
```

```
plot(nn.roc, lwd=3, col = "purple")
lines(rf.roc, lwd=3, col = "green")
lines(glm.roc, lwd=3)
legend("bottomright",title="ROC Curves",c("nnet","rf","glm"), fill=c("purple","green","black"))
```


Sensitivity
Specificity
ROC Curves
nnet
rf
glm