

Choose Your Own project: Social Determinants of Health (SDoH) as Predictors of Colon Cancer Screening

Shahidul Islam

2024-12-08

Introduction

Cancer screening plays a crucial role in the early detection, diagnosis, and primary or secondary prevention of malignancies. Social determinants of health (SDoH), such as access to adequate food and nutrition, poverty, and health insurance, along with demographic characteristics like age, gender, Body Mass Index (BMI), and racial background, can significantly influence various aspects of healthcare, including colon cancer screening. This study aimed to investigate the predictive ability of SDoH and demographic characteristics for colon cancer screening in an adult patient population.

We assessed and compared predictive ability of clinical features for colon cancer screening between the following predictive models: Logistic Regression (LR), Random Forest(RF), Support Vector Machine(SVM) and Gradient Boosting (XGB). We used “caret” package to perform 10-fold cross validation to assess the models. The metric used for assessing these model was “Area Under the Receiver Operating Characteristic (AUROC) curve.

Dataset

This project utilizes publicly available data from National Health Interview Series (NHIS, https://www.cdc.gov/nchs/nhis/documentation/2019-nhis.html#cdc_data_surveillance_section_3-using-our-data) database from 2018 to 2022. The dataset included ~176K observations, however, the colon cancer screening data along with clinically relevant features were available for ~48K subjects.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'dplyr' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
## v dplyr      1.1.4      v readr      2.1.4
```

```
## v forcats    1.0.0      v stringr    1.5.0
```

```
## v ggplot2    3.4.3      v tibble     3.2.1
```

```
## v lubridate  1.9.2      v tidyr      1.3.0
```

```
## v purrr      1.0.2
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()     masks stats::lag()
```

```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(haven)
```

```
## Warning: package 'haven' was built under R version 4.3.3
```

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##      %+%, alpha
```

```
library(rlang)
```

```
## Warning: package 'rlang' was built under R version 4.3.3
```

```
##
## Attaching package: 'rlang'
##
## The following objects are masked from 'package:purrr':
##
##      %@%, flatten, flatten_chr, flatten_dbl, flatten_int, flatten_lgl,
##      flatten_raw, invoke, splice
```

```
library(gtsummary)
```

```
## Warning: package 'gtsummary' was built under R version 4.3.3
```

```
library(glmulti)
```

```
## Warning: package 'glmulti' was built under R version 4.3.3
```

```
## Loading required package: rJava
```

```
## Warning: package 'rJava' was built under R version 4.3.2
```

```
## Loading required package: leaps
```

```
## Warning: package 'leaps' was built under R version 4.3.3
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(purrr)
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.3
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.3.3
```

```
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8
```

```
library(epiDisplay)
```

```
## Warning: package 'epiDisplay' was built under R version 4.3.3
```

```
## Loading required package: foreign
## Loading required package: survival
##
## Attaching package: 'survival'
##
## The following object is masked from 'package:caret':
##
##     cluster
##
## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:gtsummary':
##
##     select
##
## The following object is masked from 'package:dplyr':
```

```
##
##      select
##
## Loading required package: nnet
##
## Attaching package: 'epiDisplay'
##
## The following object is masked from 'package:lattice':
##
##      dotplot
##
## The following objects are masked from 'package:psych':
##
##      alpha, cs, lookup
##
## The following object is masked from 'package:ggplot2':
##
##      alpha
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following object is masked from 'package:epiDisplay':
##
##      ci
##
## The following objects are masked from 'package:stats':
##
##      cov, smooth, var
```

```
library(ResourceSelection)
```

```
## Warning: package 'ResourceSelection' was built under R version 4.3.3

## ResourceSelection 0.3-6    2023-06-27
```

```
library(survey)
```

```
## Warning: package 'survey' was built under R version 4.3.3

## Loading required package: grid
##
## Attaching package: 'survey'
##
## The following object is masked from 'package:graphics':
##
##      dotchart
```

```
library(MASS)
library(boot)
```

```
##
## Attaching package: 'boot'
##
## The following object is masked from 'package:survival':
##
##     aml
##
## The following object is masked from 'package:lattice':
##
##     melanoma
##
## The following object is masked from 'package:psych':
##
##     logit
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.3.3

## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:psych':
##
##     outlier
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(e1071) #SVM
```

```
## Warning: package 'e1071' was built under R version 4.3.3
```

```
library(xgboost)
```

```
## Warning: package 'xgboost' was built under R version 4.3.3

##
## Attaching package: 'xgboost'
##
## The following object is masked from 'package:dplyr':
##
##     slice
```

Creating analytic dataset, data wrangling and data cleaning

To better understand the data, we can print a few rows, examine its dimensions, inspect the variable types, and perform other exploratory checks.

To examine how demographic and clinical characteristics compares between different food insecurity groups, we summarize all data by food insecurity categories. Continuous variables were summarized using median (interquartile range), and categorical variables using frequency (percentages). The variables were compared using the Wilcoxon rank-sum test if the variables were continuous, and hi-square or Fisher's exact test if the variables were categorical. We used the "gtsummary" package to perform these analyses.

```
#####  
# Create analytic dataset, data wrangling and data cleaning  
#####  
  
# STEP 1. Import data -----  
setwd("C:/Users/ShahS/OneDrive/EDX.ORG/DataScienceCertificate/CAPSTONE/Choose_own_proj")  
nha<-read_sas("nha18_20_v3.sas7bdat")  
names(nha)<-tolower(names(nha))  
  
#inspecting data elements, sample data  
head(nha)
```

```
## # A tibble: 6 x 78  
##   year serial strata   psu nhishid   hhweight region pernum nhispid hhx   fmx  
##   <dbl> <dbl> <dbl> <dbl> <chr>      <dbl> <dbl> <dbl> <chr> <chr> <chr>  
## 1  2018     1   7103    19 00002018~    3338     3     1 002018~ " 00~ 01  
## 2  2018     3   7137    38 00002018~    5188     2     1 002018~ " 00~ 01  
## 3  2018     3   7137    38 00002018~    5188     2     2 002018~ " 00~ 01  
## 4  2018     4   7106    22 00002018~    5259     3     1 002018~ " 00~ 01  
## 5  2018     4   7106    22 00002018~    5259     3     2 002018~ " 00~ 01  
## 6  2018     5   7117    25 00002018~    3204     2     1 002018~ " 00~ 01  
## # i 67 more variables: px <chr>, perweight <dbl>, sampweight <dbl>,  
## #   longweight <dbl>, partweight <dbl>, fweight <dbl>, astatflg <dbl>,  
## #   cstatflg <dbl>, age <dbl>, sex <dbl>, racenew <dbl>, hispeth <dbl>,  
## #   gotstampfam <dbl>, poverty <dbl>, fsrawscore <dbl>, fsstatdet <dbl>,  
## #   health <dbl>, weight <dbl>, bmecat <dbl>, hiprivate <dbl>, histate <dbl>,  
## #   hichip <dbl>, himcaid <dbl>, himcare <dbl>, hiothgov <dbl>, hinone <dbl>,  
## #   cnbres <dbl>, cncerv <dbl>, cncoln <dbl>, cnpros <dbl>, bsthev <dbl>, ...
```

```
names(nha)
```

```
## [1] "year"           "serial"         "strata"  
## [4] "psu"            "nhishid"        "hhweight"  
## [7] "region"         "pernum"         "nhispid"  
## [10] "hhx"            "fmx"            "px"  
## [13] "perweight"      "sampweight"     "longweight"  
## [16] "partweight"     "fweight"        "astatflg"  
## [19] "cstatflg"       "age"            "sex"  
## [22] "racenew"        "hispeth"        "gotstampfam"  
## [25] "poverty"        "fsrawscore"     "fsstatdet"  
## [28] "health"         "weight"         "bmecat"  
## [31] "hiprivate"      "histate"        "hichip"  
## [34] "himcaid"        "himcare"        "hiothgov"
```

```
## [37] "hinone"           "cnbres"           "cncerv"
## [40] "cncoln"           "cnpros"           "bsthev"
## [43] "colgev"           "colsigev"         "mamev"
## [46] "crvtstev"         "papev"            "psaev"
## [49] "food_insecure"    "food_insecure_r"  "female"
## [52] "racenew_r"        "white"            "black"
## [55] "asian"            "other_race"       "hispanic"
## [58] "foodstamp"        "food_security_score" "private_ins"
## [61] "state_ins"        "chi_ins"          "medicaid"
## [64] "medicare"         "other_gov_ins"    "no_ins"
## [67] "insurance"        "blood_stool_test" "cologuard"
## [70] "colonscopy"       "colon_scr"        "mammogram"
## [73] "pap_test"         "cervc_test"       "cervical_scr"
## [76] "prostate_scr"     "health_r"         "poverty_r"
```

```
str(nha)
```

```
## tibble [176,562 x 78] (S3: tbl_df/tbl/data.frame)
## $ year           : num [1:176562] 2018 2018 2018 2018 2018 ...
## $ serial         : num [1:176562] 1 3 3 4 4 5 5 6 7 7 ...
## $ strata         : num [1:176562] 7103 7137 7137 7106 7106 ...
## $ psu            : num [1:176562] 19 38 38 22 22 25 25 34 152 152 ...
## $ nhishid        : chr [1:176562] "000020180000001" "000020180000004" "000020180000004" "000020180000004" ...
## $ hhweight       : num [1:176562] 3338 5188 5188 5259 5259 ...
## $ region         : num [1:176562] 3 2 2 3 3 2 2 3 1 1 ...
## $ pernum         : num [1:176562] 1 1 2 1 2 1 2 1 1 2 ...
## $ nhispid        : chr [1:176562] "00201800000010101" "00201800000040101" "00201800000040102" "00201800000040102" ...
## $ hhx            : chr [1:176562] " 000001" " 000004" " 000004" " 000006" ...
## $ fmx            : chr [1:176562] "01" "01" "01" "01" ...
## $ px            : chr [1:176562] "01" "01" "02" "01" ...
## $ perweight      : num [1:176562] 3539 5656 5848 6009 5738 ...
## $ sampweight     : num [1:176562] 3915 0 0 16978 0 ...
## $ longweight     : num [1:176562] NA NA NA NA NA NA NA NA NA ...
## $ partweight     : num [1:176562] NA NA NA NA NA NA NA NA NA ...
## $ fweight        : num [1:176562] 3539 5656 5656 5415 5415 ...
## $ astatflg       : num [1:176562] 1 3 2 1 3 1 3 1 1 3 ...
## $ ...            : chr [1:176562] "Sample adult flag"
```

```

## $ cstatflg      : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Sample child flag"
## $ age           : num [1:176562] 79 55 49 37 36 29 34 75 39 33 ...
##   ..- attr(*, "label")= chr "Age"
## $ sex           : num [1:176562] 2 2 1 1 2 1 2 1 1 2 ...
##   ..- attr(*, "label")= chr "Sex"
## $ racenew       : num [1:176562] 100 100 100 100 100 100 100 400 100 100 ...
##   ..- attr(*, "label")= chr "Self-reported Race (Post-1997 OMB standards)"
## $ hispeth       : num [1:176562] 10 10 10 10 10 10 10 10 10 10 ...
##   ..- attr(*, "label")= chr "Hispanic ethnicity"
## $ gotstampfam   : num [1:176562] 10 10 10 10 10 10 10 21 10 10 ...
##   ..- attr(*, "label")= chr "Any family member received food stamp/SNAP benefits last calendar year"
## $ poverty       : num [1:176562] 22 24 24 37 37 36 36 22 34 34 ...
##   ..- attr(*, "label")= chr "Ratio of family income to poverty threshold"
## $ fsrawscore    : num [1:176562] 0 2 2 0 0 0 0 8 0 0 ...
##   ..- attr(*, "label")= chr "Family's raw score on the 30-day food security scale"
## $ fsstatdet     : num [1:176562] 1 2 2 1 1 1 1 4 1 1 ...
##   ..- attr(*, "label")= chr "Detailed family-level food security status on the 30-day food security"
## $ health        : num [1:176562] 3 3 3 3 2 2 1 5 2 2 ...
##   ..- attr(*, "label")= chr "Health status"
## $ weight        : num [1:176562] 129 0 0 235 0 996 0 138 170 0 ...
##   ..- attr(*, "label")= chr "Weight in pounds without shoes"
## $ bmicat        : num [1:176562] 2 NA NA 4 NA 4 NA 2 2 NA ...
##   ..- attr(*, "label")= chr "Categorical body mass index"
## $ hiprivate     : num [1:176562] 1 2 2 2 2 2 2 2 2 2 ...
##   ..- attr(*, "label")= chr "Has any private health insurance"
## $ histate       : num [1:176562] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Has state-sponsored health plan insurance"
## $ hichip        : num [1:176562] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Has Children's Health Insurance Program"
## $ himcaid       : num [1:176562] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Has Medicaid insurance"
## $ himcare       : num [1:176562] 2 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Has Medicare insurance"
## $ hiothgov      : num [1:176562] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Has other government program insurance"
## $ hinone        : num [1:176562] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Has no health insurance"
## $ cnbres        : num [1:176562] 0 0 0 1 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Ever had cancer: Breast"
## $ cncerv        : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Ever had cancer: Cervix"
## $ cncoln       : num [1:176562] 0 0 0 1 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Ever had cancer: Colon"
## $ cnpros        : num [1:176562] 0 0 0 1 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Ever had cancer: Prostate"
## $ bsthev        : num [1:176562] 1 0 0 0 0 0 0 2 0 0 ...
##   ..- attr(*, "label")= chr "Ever had blood stool test using home test kit"
## $ colgev        : num [1:176562] 1 0 0 0 0 0 0 1 0 0 ...
##   ..- attr(*, "label")= chr "Ever had Cologuard test"
## $ colsigev      : num [1:176562] 1 0 0 0 0 0 0 1 0 0 ...
##   ..- attr(*, "label")= chr "Ever had colonoscopy, sigmoidoscopy, or both"
## $ mamev         : num [1:176562] 2 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Ever had a mammogram"

```



```
## $ crvtstev      : num [1:176562] NA NA NA NA NA NA NA NA NA NA NA ...
##   ..- attr(*, "label")= chr "Ever had cervical cancer test"
## $ papev         : num [1:176562] 2 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "label")= chr "Ever had Pap smear test"
## $ psaev         : num [1:176562] 0 0 0 0 0 0 0 0 1 0 0 ...
##   ..- attr(*, "label")= chr "Ever had a PSA test"
## $ food_insecure : num [1:176562] 0 1 1 0 0 0 0 1 0 0 ...
## $ food_insecure_r : num [1:176562] 1 2 2 1 1 1 1 3 1 1 ...
## $ female         : num [1:176562] 1 1 0 0 1 0 1 0 0 1 ...
## $ racenew_r      : chr [1:176562] "White" "White" "White" "White" ...
## $ white          : num [1:176562] 1 1 1 1 1 1 1 1 0 1 1 ...
## $ black          : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
## $ asian          : num [1:176562] 0 0 0 0 0 0 0 0 1 0 0 ...
## $ other_race     : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
## $ hispanic       : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
## $ foodstamp      : num [1:176562] 0 0 0 0 0 0 0 0 1 0 0 ...
## $ food_security_score: num [1:176562] 0 2 2 0 0 0 0 8 0 0 ...
## $ private_ins    : num [1:176562] 0 1 1 1 1 1 1 1 1 1 ...
## $ state_ins      : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
## $ chi_ins        : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
## $ medicaid      : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
## $ medicare       : num [1:176562] 1 0 0 0 0 0 0 0 0 0 ...
## $ other_gov_ins  : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
## $ no_ins         : num [1:176562] 0 0 0 0 0 0 0 0 0 0 ...
## $ insurance      : num [1:176562] 2 1 1 1 1 1 1 1 1 1 ...
## $ blood_stool_test : num [1:176562] 0 NA NA NA NA NA NA 1 NA NA ...
## $ cologuard      : num [1:176562] 0 NA NA NA NA NA NA 0 NA NA ...
## $ colonoscopy    : num [1:176562] 1 NA NA NA NA NA NA 1 NA NA ...
## $ colon_scr      : num [1:176562] 1 NA NA NA NA NA NA 1 NA NA ...
## $ mammogram      : num [1:176562] 1 NA NA NA NA NA NA NA NA NA ...
## $ pap_test       : num [1:176562] 1 NA NA NA NA NA NA NA NA NA ...
## $ cervc_test     : num [1:176562] NA NA NA NA NA NA NA NA NA ...
## $ cervical_scr   : num [1:176562] 1 NA NA NA NA NA NA NA NA NA ...
## $ prostate_scr   : num [1:176562] NA NA NA NA NA NA NA 0 NA NA ...
## $ health_r       : chr [1:176562] "Good" "Good" "Good" "Good" ...
## $ poverty_r      : chr [1:176562] "1-1.99" "1-1.99" "1-1.99" ">=5" ...
```

```
#recoding missing values to NA
```

```
nha1<-nha%>%
```

```
  dplyr::mutate(
    racenew_r1=ifelse(racenew_r=='', NA, racenew_r),
    health_r1=ifelse(health_r=='', NA, health_r),
    bmicat=ifelse(bmicat=='', NA, bmicat),
    poverty_r=ifelse(poverty_r=='', NA, poverty_r)
  )
```

```
#selecting appropriate variables for the summary table
```

```
nha2<- nha1%>%dplyr::select(food_insecure, food_insecure_r, year, age, bmicat, female, racenew_r1, white)
```

```
#Summary of demographics and clinical characteristics by group
```

```
nha2%>%
```

```
  tbl_summary(by = food_insecure_r, missing = "no") %>%
  add_p(pvalue_fun = ~style_pvalue(.x, digits = 3)) %>%
```

```
add_overall() %>%
add_n() %>%
bold_labels()
```

207 missing rows in the "food_insecure_r" column have been removed.

dataset creation for model building

We decided to predict “colon cancer screening” using relevant clinical features and different modeling approach. First we need to create a complete case dataset using the clinically relevant features.

```
# Select variables for colon cancer screening
nha3 <- nha2 %>%
  dplyr::select(
    food_insecure_r, year, age, bmicat, female, racenew_r1, insurance, poverty_r, colon_scr
  )

# Ensure variables are factors
factor_vars <- c("food_insecure_r", "racenew_r1", "female", "insurance", "bmicat", "poverty_r")
nha3[factor_vars] <- lapply(nha3[factor_vars], factor)

# Relevel factors
nha3$food_insecure_r <- relevel(nha3$food_insecure_r, ref = "1")
nha3$racenew_r1 <- relevel(nha3$racenew_r1, ref = "White")
nha3$female <- relevel(nha3$female, ref = "0")
nha3$insurance <- relevel(nha3$insurance, ref = "1")
nha3$bmicat <- relevel(nha3$bmicat, ref = "2")
nha3$poverty_r <- relevel(nha3$poverty_r, ref = "<1")

# Remove rows with missing values, N=48676
nha3_colon_c <- nha3[complete.cases(nha3), ]
```

####Model1. Logistic Regression#####

First, we tried simpler Logistic Regression Model to predict colon cancer screening.

```
model_colon=glm(colon_scr~as.factor(food_insecure_r)+age+as.factor(bmicat)+as.factor(female)+as.factor(
  as.factor(poverty_r)+as.factor(insurance), family=binomial(link=logit), data = nha3_

summary(model_colon)
```

```
##
## Call:
## glm(formula = colon_scr ~ as.factor(food_insecure_r) + age +
##      as.factor(bmicat) + as.factor(female) + as.factor(racenew_r1) +
##      as.factor(poverty_r) + as.factor(insurance), family = binomial(link = logit),
##      data = nha3_colon_c)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -5.289979 0.087659 -60.347 < 2e-16 ***
## as.factor(food_insecure_r)2 -0.034279 0.056133 -0.611 0.541412
## as.factor(food_insecure_r)3 0.256745 0.048767 5.265 1.4e-07 ***
## age 0.099403 0.001259 78.977 < 2e-16 ***
## as.factor(bmicat)1 -0.396075 0.104218 -3.800 0.000144 ***
## as.factor(bmicat)3 0.082349 0.030392 2.710 0.006738 **
## as.factor(bmicat)4 0.257374 0.030979 8.308 < 2e-16 ***
## as.factor(female)1 0.064907 0.024376 2.663 0.007752 **
## as.factor(racenew_r1)Asian -0.463466 0.051730 -8.959 < 2e-16 ***
## as.factor(racenew_r1)Black 0.089330 0.038092 2.345 0.019022 *
## as.factor(racenew_r1)Other -0.051246 0.069117 -0.741 0.458427
## as.factor(poverty_r)>=5 0.803749 0.049604 16.203 < 2e-16 ***
## as.factor(poverty_r)1-1.99 0.020564 0.047564 0.432 0.665497
## as.factor(poverty_r)2-4.99 0.404282 0.046040 8.781 < 2e-16 ***
## as.factor(insurance)2 -0.047510 0.032846 -1.446 0.148053
## as.factor(insurance)3 -0.584736 0.042299 -13.824 < 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: 54037 on 48675 degrees of freedom
## Residual deviance: 43533 on 48660 degrees of freedom
## AIC: 43565
##
## Number of Fisher Scoring iterations: 5
```

```
# Predict probabilities for each observation
predicted_probs <- predict(model_colon, nha3_colon_c, type = "response")
# Calculate AUC--model discrimination
roc_obj_logit<-roc(nha3_colon_c$colon_scr, predicted_probs)
```

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

```
## Setting direction: controls < cases
```

```
auc_logistic <-as.numeric(roc(nha3_colon_c$colon_scr, predicted_probs)$auc)
```

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

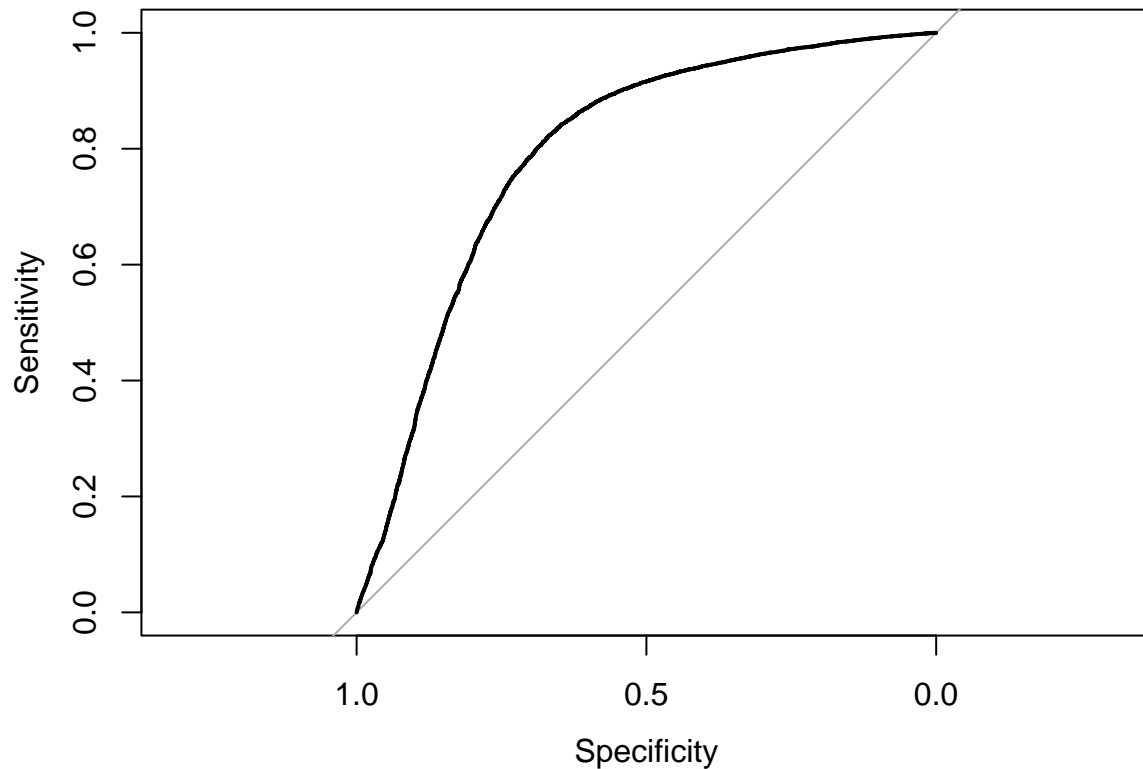
```
## Setting direction: controls < cases
```

```
auc_logistic
```

```
## [1] 0.7918455
```

```
plot(roc_obj_logit, main = "ROC Curve for Logistic Regression Model")
```

ROC Curve for Logistic Regression Model



```
#goodness of fit test for logistic regression
complete_cases <- complete.cases(nha3_colon_c$colon_scr, predicted_probs)
observed <- nha3_colon_c$colon_scr[complete_cases]
predicted_probs <- predicted_probs[complete_cases]
hoslem.test(observed, predicted_probs, g=10)
```

```
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: observed, predicted_probs
## X-squared = 1554.2, df = 8, p-value < 2.2e-16
```

```
#Confusion Matrix and Accuracy:
predicted_probs1 <- predict(model_colon, nha3_colon_c, type = "response")
predicted <- ifelse(predicted_probs1 > 0.5, 1, 0)
conf_matrix <- table(predicted, nha3_colon_c$colon_scr)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
sensitivity <- conf_matrix[2, 2] / sum(conf_matrix[2, ])
specificity <- conf_matrix[1, 1] / sum(conf_matrix[1, ])
accuracy
```

```
## [1] 0.8108924
```

```
sensitivity
```

```
## [1] 0.8317757
```

```
specificity
```

```
## [1] 0.6876506
```

```
###cross validation using caret package for logistic regression model
```

```
# Convert the outcome of interest "colon_scr" var to a factor with valid levels
```

```
#for cross validation using caret package
```

```
nha3_colon_c$colon_scr_r <- factor(  
  nha3_colon_c$colon_scr,  
  levels = c(0, 1),          # Ensure correct levels  
  labels = c("No", "Yes")    # Use valid string labels  
)
```

```
# Define the cross-validation setup--we will use this setup for all subsequent models for cross-validat
```

```
train_control <- trainControl(  
  method = "cv",  
  number = 10,  
  summaryFunction = twoClassSummary,  
  classProbs = TRUE # Enables probability predictions  
)
```

```
#####cross validation for Logistic Regression model#####
```

```
cv_model <- train(  
  colon_scr_r ~ as.factor(food_insecure_r) + age + as.factor(bmicat) +  
    as.factor(female) + as.factor(racenew_r1) +  
    as.factor(poverty_r) + as.factor(insurance),  
  data = nha3_colon_c,  
  method = "glm",  
  family = "binomial", # Specify logistic regression  
  trControl = train_control,  
  metric = "ROC" # Use ROC as the evaluation metric  
)
```

```
# Print the model summary
```

```
print(cv_model)
```

```
## Generalized Linear Model
```

```
##
```

```
## 48676 samples
```

```
## 7 predictor
```

```
## 2 classes: 'No', 'Yes'
```

```
##
```

```
## No pre-processing
```

```
## Resampling: Cross-Validated (10 fold)
```

```
## Summary of sample sizes: 43808, 43808, 43809, 43809, 43808, 43809, ...
```

```
## Resampling results:
```

```
##
## ROC      Sens      Spec
## 0.791549 0.4092125 0.9404739
```

```
summary(cv_model)
```

```
##
## Call:
## NULL
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.289979   0.087659  -60.347  < 2e-16 ***
## 'as.factor(food_insecure_r)2' -0.034279   0.056133  -0.611  0.541412
## 'as.factor(food_insecure_r)3'  0.256745   0.048767   5.265  1.4e-07 ***
## age              0.099403   0.001259  78.977  < 2e-16 ***
## 'as.factor(bmicat)1'    -0.396075   0.104218  -3.800  0.000144 ***
## 'as.factor(bmicat)3'     0.082349   0.030392   2.710  0.006738 **
## 'as.factor(bmicat)4'     0.257374   0.030979   8.308  < 2e-16 ***
## 'as.factor(female)1'     0.064907   0.024376   2.663  0.007752 **
## 'as.factor(racenew_r1)Asian' -0.463466   0.051730  -8.959  < 2e-16 ***
## 'as.factor(racenew_r1)Black'  0.089330   0.038092   2.345  0.019022 *
## 'as.factor(racenew_r1)Other' -0.051246   0.069117  -0.741  0.458427
## 'as.factor(poverty_r)>=5'    0.803749   0.049604  16.203  < 2e-16 ***
## 'as.factor(poverty_r)1-1.99'  0.020564   0.047564   0.432  0.665497
## 'as.factor(poverty_r)2-4.99'  0.404282   0.046040   8.781  < 2e-16 ***
## 'as.factor(insurance)2'    -0.047510   0.032846  -1.446  0.148053
## 'as.factor(insurance)3'    -0.584736   0.042299 -13.824  < 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: 54037  on 48675  degrees of freedom
## Residual deviance: 43533  on 48660  degrees of freedom
## AIC: 43565
##
## Number of Fisher Scoring iterations: 5
```

```
extract_auc_logistic=as.data.frame(cv_model$resample)
auc_logistic_xvalidated<-as.numeric(mean(extract_auc_logistic$ROC), na.rm=TRUE)

#saving AUC estimates in a dataframe
auc_table <- as.data.frame(tibble(Model ="Logistic Regression model",
                                   AUC_single_model=auc_logistic, AUC_10F_xvalidated = auc_logistic_xval.
auc_table %>% knitr::kable()
```

Model	AUC_single_model	AUC_10F_xvalidated
Logistic Regression model	0.7918455	0.791549

```
auc_table
```

```
##                               Model AUC_single_model AUC_10F_xvalidated
## 1 Logistic Regression model      0.7918455           0.791549
```

Model2. Random Forest Model

We then implemented Random Forest Model to predict colon cancer screening.

```
nha3_colon_c$colon_scr <- as.factor(nha3_colon_c$colon_scr)
```

```
# Fit a random forest model
```

```
model_rf <- randomForest(colon_scr_r ~ food_insecure_r + age +bmicat +
  female + racenew_r1 +poverty_r + insurance, data = nha3_colon_c,
  ntree = 100, importance = TRUE)
```

```
# View model summary
```

```
print(model_rf)
```

```
##
```

```
## Call:
```

```
## randomForest(formula = colon_scr_r ~ food_insecure_r + age +          bmicat + female + racenew_r1 + po
```

```
##               Type of random forest: classification
```

```
##               Number of trees: 100
```

```
## No. of variables tried at each split: 2
```

```
##
```

```
##               OOB estimate of  error rate: 17.79%
```

```
## Confusion matrix:
```

```
##           No   Yes class.error
```

```
## No  6190  5662  0.47772528
```

```
## Yes 2997 33827  0.08138714
```

```
# Feature importance
```

```
importance(model_rf)
```

```
##               No           Yes MeanDecreaseAccuracy MeanDecreaseGini
```

```
## food_insecure_r 2.0679647 7.718655           8.863877       131.87823
```

```
## age             62.5411473 56.688409           63.813348       4504.67864
```

```
## bmicat          -2.7383763 9.422010           8.609955        188.11888
```

```
## female          3.2801213 4.589399           6.523065        76.85793
```

```
## racenew_r1      0.2420003 6.569755           5.331814       201.78889
```

```
## poverty_r       8.3163804 6.083922           16.620182       266.09096
```

```
## insurance       8.9919493 6.841323           20.996893       411.32422
```

```
# Predict probabilities
```

```
pred_probs <- predict(model_rf, nha3_colon_c, type = "prob")[, 2] # Probability of class 1
```

```
# Compute AUC
```

```
roc_obj_rf <- roc(nha3_colon_c$colon_scr, pred_probs)
```

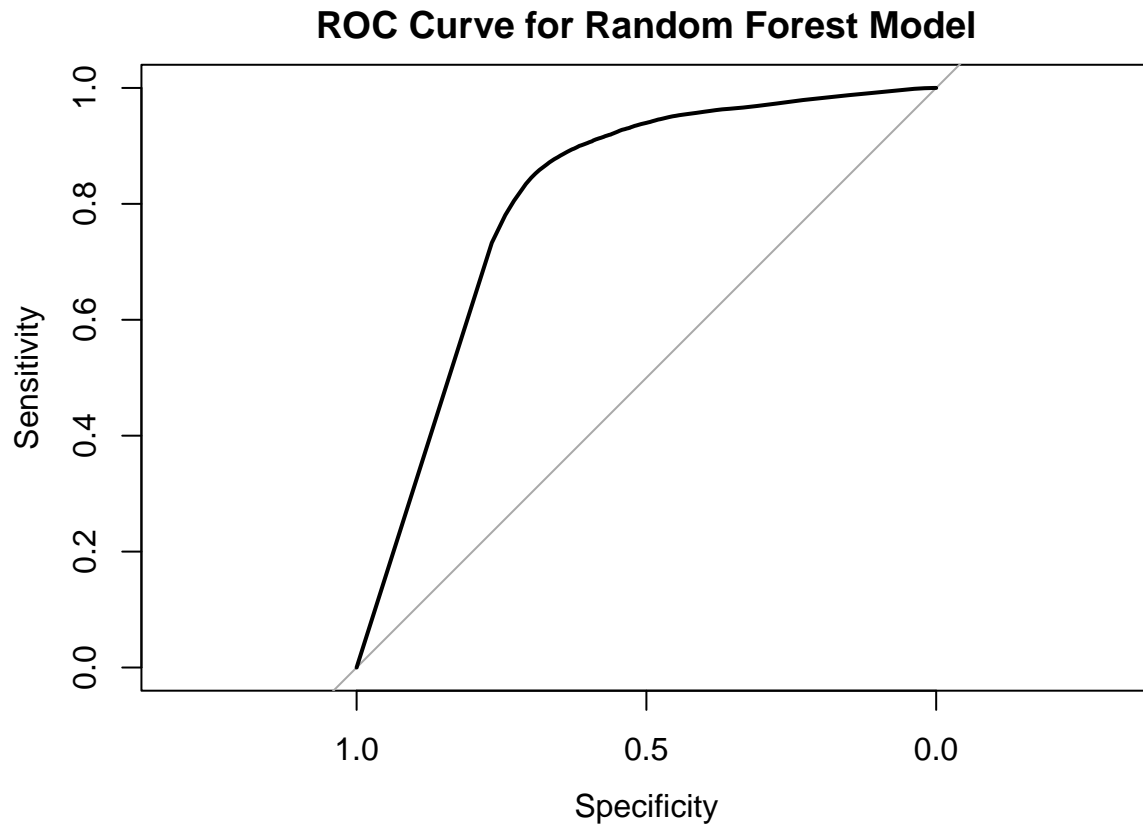
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
auc_rf <- auc(roc_obj_rf)
print(auc_rf)
```

```
## Area under the curve: 0.8065
```

```
plot(roc_obj_rf, main = "ROC Curve for Random Forest Model")
```



```
#####cross validation for random forest model#####
```

```
# Train the random forest model using caret
```

```
cv_model_rf <- train(
  colon_scr_r ~ food_insecure_r + age + bmicat +
    female + racenew_r1 +
    poverty_r + insurance,
  data = nha3_colon_c,
  method = "rf",                # Specify Random Forest as the model
  trControl = train_control,    # Use cross-validation setup that was previously defined (see above)
  metric = "ROC",               # Optimize based on ROC AUC
  ntree = 100,                  # Number of trees
  importance = TRUE              # Track variable importance
)
```



```
# Print the model summary and plot
print(cv_model_rf)
```

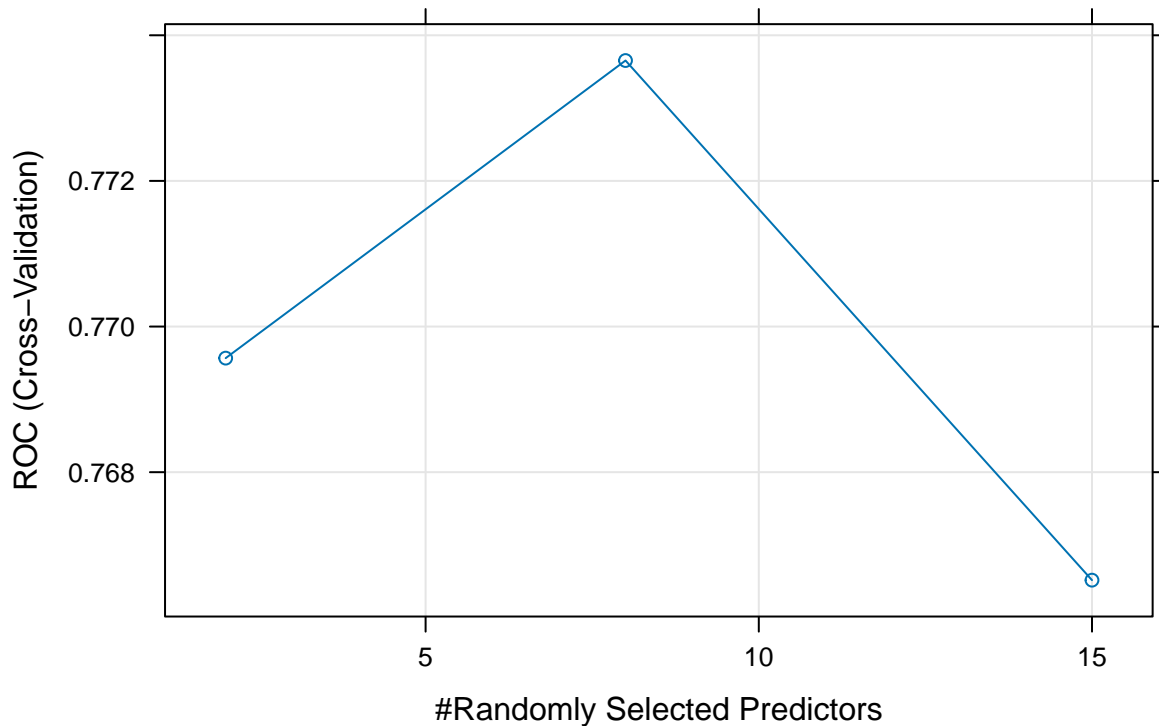
```
## Random Forest
##
## 48676 samples
##    7 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 43808, 43808, 43808, 43809, 43809, 43808, ...
## Resampling results across tuning parameters:
##
##  mtry  ROC          Sens          Spec
##    2   0.7695661  0.4924013  0.9267047
##    8   0.7736533  0.5080158  0.9103571
##   15   0.7665172  0.5084372  0.8950682
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 8.
```

```
summary(cv_model_rf)
```

```
##           Length Class      Mode
## call           6 -none-    call
## type           1 -none- character
## predicted     48676 factor   numeric
## err.rate       300 -none-    numeric
## confusion       6 -none-    numeric
## votes         97352 matrix   numeric
## oob.times      48676 -none-    numeric
## classes        2 -none-    character
## importance      60 -none-    numeric
## importanceSD    45 -none-    numeric
## localImportance 0 -none-    NULL
## proximity       0 -none-    NULL
## ntree          1 -none-    numeric
## mtry           1 -none-    numeric
## forest         14 -none-    list
## y             48676 factor   numeric
## test           0 -none-    NULL
## inbag           0 -none-    NULL
## xNames         15 -none-    character
## problemType     1 -none-    character
## tuneValue       1 data.frame list
## obsLevels       2 -none-    character
## param          2 -none-    list
```

```
plot(cv_model_rf, main = "Cross-Validated ROC vs. # of predictors")
```

Cross-Validated ROC vs. # of predictors



```
extract_auc_rf=as.data.frame(cv_model_rf$resample)
auc_rf_xvalidated<-mean(extract_auc_rf$ROC, na.rm=TRUE)

auc_rf <- as.numeric(auc_rf)
auc_rf_xvalidated <- as.numeric(auc_rf_xvalidated)

#adding AUC to the dataframe
auc_table <- auc_table %>% add_row(Model="Random Forest Model", AUC_single_model=auc_rf,
                                   AUC_10F_xvalidated = auc_rf_xvalidated)
auc_table
```

```
##               Model AUC_single_model AUC_10F_xvalidated
## 1 Logistic Regression model      0.7918455      0.7915490
## 2      Random Forest Model      0.8064965      0.7736533
```

Model3. Support Vector Machine (SVM) Model

Cross validated AUROC for random forest model was worse than Logistic Regression Model. So, we tested if the Support Vector Machine (SVM) model perform better.

```
model_svm <- svm(colon_scr ~ food_insecure_r + age +bmicat +
                 female + racenew_r1 +
                 poverty_r + insurance,
                 data = nha3_colon_c, kernel = "radial", probability = TRUE)
```

```
summary(model_svm)
```

```
##  
## Call:  
## svm(formula = colon_scr ~ food_insecure_r + age + bmicat + female +  
##       racenew_r1 + poverty_r + insurance, data = nha3_colon_c, kernel = "radial",  
##       probability = TRUE)  
##  
##  
## Parameters:  
##   SVM-Type:  C-classification  
##   SVM-Kernel: radial  
##         cost:  1  
##  
## Number of Support Vectors:  19883  
##  
## ( 10036 9847 )  
##  
##  
## Number of Classes:  2  
##  
## Levels:  
##  0 1
```

```
pred <- predict(model_svm, nha3_colon_c, probability = TRUE)  
probabilities <- attr(pred, "probabilities")[, 2] # Probabilities for the positive class  
roc_obj_svm <- roc(nha3_colon_c$colon_scr, probabilities)
```

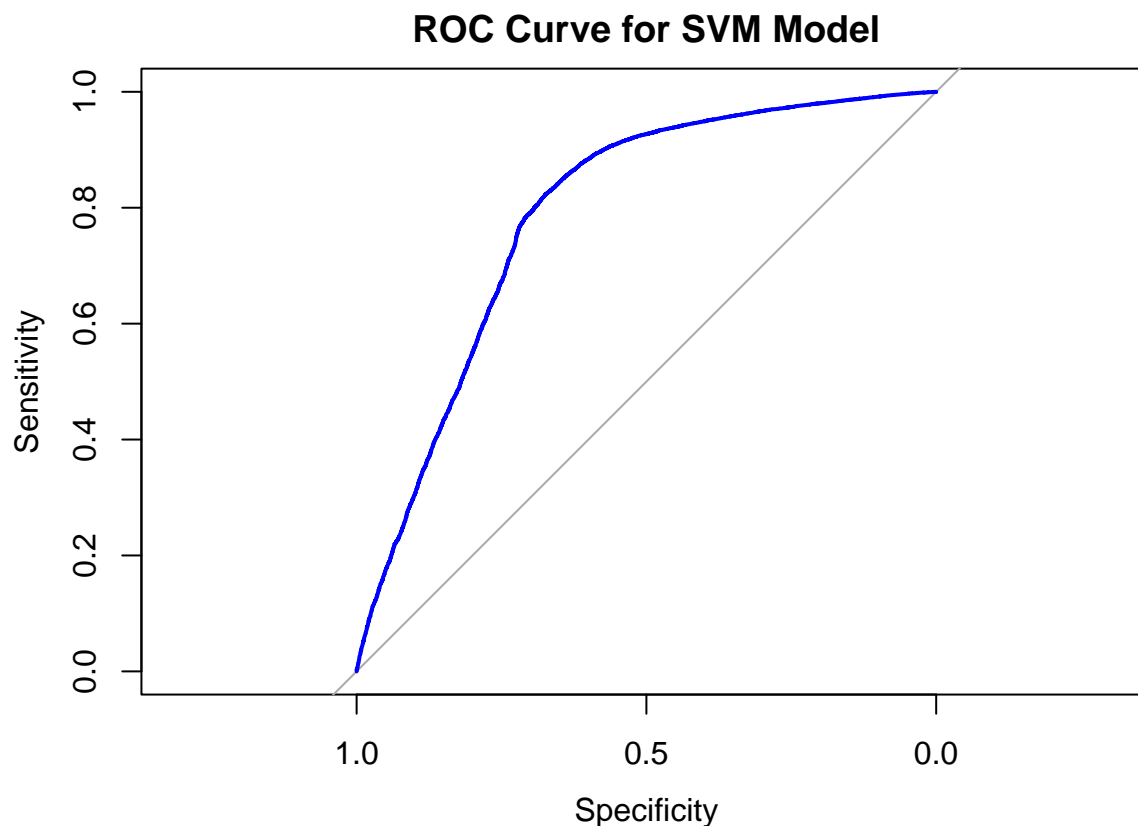
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls > cases
```

```
auc_value_svm <- as.numeric(auc(roc_obj_svm))  
print(auc_value_svm)
```

```
## [1] 0.7880106
```

```
plot(roc_obj_svm, main = "ROC Curve for SVM Model", col = "blue", lwd = 2)
```



#####cross validation for support vector machine (SVM) model#####

Train the SVM model using caret

```
cv_model_svm <- train(
  colon_scr_r ~ food_insecure_r + age + bmicat +
    female + racenew_r1 +
    poverty_r + insurance,
  data = nha3_colon_c,
  method = "svmRadial",      # Specify the radial kernel SVM
  trControl = train_control, # Use cross-validation setup
  metric = "ROC",            # Optimize based on ROC AUC
  probability = TRUE         # Enable probability predictions (needed for AUC)
)
```

```
## line search fails -1.716433 0.02871571 1.027305e-05 5.186023e-06 -2.653176e-09 8.363816e-10 -2.291877
```

```
## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class prediction calculations failed; returning NAs
```

```
## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class probability calculations failed; returning NAs
```

```
## Warning in data.frame(..., check.names = FALSE): row names were found from a
## short variable and have been discarded
```

```

## line search fails -1.710229 0.02103636 1.083968e-05 5.361669e-06 -2.798152e-09 8.898904e-10 -2.555977

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a
## short variable and have been discarded

## line search fails -1.722203 0.03080995 1.072897e-05 5.161338e-06 -2.839234e-09 9.557544e-10 -2.552909

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a
## short variable and have been discarded

## line search fails -1.716274 0.02991708 1.00915e-05 4.92156e-06 -2.646507e-09 8.792505e-10 -2.237994e-09

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a
## short variable and have been discarded

## line search fails -1.715064 0.01894875 1.179297e-05 5.996636e-06 -3.009391e-09 9.068348e-10 -3.005177

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a
## short variable and have been discarded

## line search fails -1.722878 0.02988081 1.200309e-05 6.221355e-06 -3.050128e-09 9.039943e-10 -3.098699

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class prediction calculations failed; returning NAs

```

```
## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a
## short variable and have been discarded

## line search fails -1.719029 0.03011231 1.010558e-05 4.853659e-06 -2.677458e-09 9.078303e-10 -2.26509

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a
## short variable and have been discarded

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## line search fails -1.718095 0.02265773 1.107673e-05 5.224213e-06 -2.659414e-09 9.128664e-10 -2.46885
```

```
# Print the model summary
print(cv_model_svm)
```

```
## Support Vector Machines with Radial Basis Function Kernel
##
## 48676 samples
##      7 predictor
##      2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 43809, 43807, 43809, 43809, 43809, 43807, ...
## Resampling results across tuning parameters:
##
##      C      ROC      Sens      Spec
##  0.25  0.8033707  0.4329512  0.9464205
##  0.50  0.7896869  0.4186619  0.9438407
##  1.00  0.7846907  0.4022931  0.9472082
##
## Tuning parameter 'sigma' was held constant at a value of 0.06345747
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were sigma = 0.06345747 and C = 0.25.
```

```
extract_auc_svm=as.data.frame(cv_model_svm$resample)

extract_auc_svm$ROC=as.numeric(extract_auc_svm$ROC)

auc_svm_xvalidated<-mean(extract_auc_svm$ROC, na.rm=TRUE)
```

```
#adding AUC to the dataframe
```

```
auc_table <- auc_table %>% add_row(Model="Support Vector Machine model", AUC_single_model=auc_value_svm,  
                                   AUC_10F_xvalidated = auc_svm_xvalidated)  
auc_table
```

```
##                               Model AUC_single_model AUC_10F_xvalidated  
## 1   Logistic Regression model      0.7918455      0.7915490  
## 2   Random Forest Model           0.8064965      0.7736533  
## 3 Support Vector Machine model     0.7880106      0.8033707
```

Model4. Gradient Boosting (i.e., XGBoost)

The SVM model performed slightly better than Logistic regression model, and of course better than Random Forest Model. So, we finally built Gradient Boosting Model.

```
# Prepare data for XGBoost
```

```
nha3_colon_c$colon_scr <- as.numeric(as.character(nha3_colon_c$colon_scr))
```

```
X <- model.matrix(colon_scr ~ food_insecure_r + age +bmicat +  
                  female + racenew_r1 +  
                  poverty_r + insurance,  
                  data = nha3_colon_c)[,-1]  
y <- nha3_colon_c$colon_scr
```

```
# Train-test split
```

```
set.seed(1157)  
train_idx <- sample(1:nrow(nha3_colon_c), 0.8 * nrow(nha3_colon_c))
```

```
dtrain <- xgb.DMatrix(data = X[train_idx,], label = y[train_idx])  
dtest  <- xgb.DMatrix(data = X[-train_idx,], label = y[-train_idx])
```

```
# dtrain_labels <- as.numeric(as.factor(your_labels)) - 1
```

```
# dtrain <- xgb.DMatrix(data = your_features, label = dtrain_labels)
```

```
# Train XGBoost
```

```
model_xgb <- xgboost(data = dtrain, objective = "binary:logistic", nrounds = 100)
```

```
## [1] train-logloss:0.573017  
## [2] train-logloss:0.509275  
## [3] train-logloss:0.472552  
## [4] train-logloss:0.450502  
## [5] train-logloss:0.437280  
## [6] train-logloss:0.429271  
## [7] train-logloss:0.424180  
## [8] train-logloss:0.420818  
## [9] train-logloss:0.418873  
## [10] train-logloss:0.417329  
## [11] train-logloss:0.416254  
## [12] train-logloss:0.415350  
## [13] train-logloss:0.414684
```

```
## [14] train-logloss:0.414054
## [15] train-logloss:0.413443
## [16] train-logloss:0.412816
## [17] train-logloss:0.412310
## [18] train-logloss:0.411885
## [19] train-logloss:0.411391
## [20] train-logloss:0.410890
## [21] train-logloss:0.410491
## [22] train-logloss:0.410043
## [23] train-logloss:0.409669
## [24] train-logloss:0.409123
## [25] train-logloss:0.408747
## [26] train-logloss:0.408301
## [27] train-logloss:0.407959
## [28] train-logloss:0.407583
## [29] train-logloss:0.407367
## [30] train-logloss:0.407030
## [31] train-logloss:0.406823
## [32] train-logloss:0.406492
## [33] train-logloss:0.406263
## [34] train-logloss:0.406063
## [35] train-logloss:0.405552
## [36] train-logloss:0.405317
## [37] train-logloss:0.404856
## [38] train-logloss:0.404546
## [39] train-logloss:0.404073
## [40] train-logloss:0.403851
## [41] train-logloss:0.403651
## [42] train-logloss:0.403373
## [43] train-logloss:0.403147
## [44] train-logloss:0.402864
## [45] train-logloss:0.402707
## [46] train-logloss:0.402480
## [47] train-logloss:0.402116
## [48] train-logloss:0.401810
## [49] train-logloss:0.401636
## [50] train-logloss:0.401429
## [51] train-logloss:0.401079
## [52] train-logloss:0.400715
## [53] train-logloss:0.400605
## [54] train-logloss:0.400416
## [55] train-logloss:0.400058
## [56] train-logloss:0.399846
## [57] train-logloss:0.399644
## [58] train-logloss:0.399417
## [59] train-logloss:0.399239
## [60] train-logloss:0.399057
## [61] train-logloss:0.398758
## [62] train-logloss:0.398404
## [63] train-logloss:0.398192
## [64] train-logloss:0.398024
## [65] train-logloss:0.397814
## [66] train-logloss:0.397667
## [67] train-logloss:0.397408
```



```
## [68] train-logloss:0.397176
## [69] train-logloss:0.396986
## [70] train-logloss:0.396818
## [71] train-logloss:0.396727
## [72] train-logloss:0.396575
## [73] train-logloss:0.396241
## [74] train-logloss:0.396044
## [75] train-logloss:0.395763
## [76] train-logloss:0.395476
## [77] train-logloss:0.395257
## [78] train-logloss:0.395024
## [79] train-logloss:0.394780
## [80] train-logloss:0.394631
## [81] train-logloss:0.394288
## [82] train-logloss:0.394170
## [83] train-logloss:0.394001
## [84] train-logloss:0.393866
## [85] train-logloss:0.393712
## [86] train-logloss:0.393463
## [87] train-logloss:0.393303
## [88] train-logloss:0.393102
## [89] train-logloss:0.393014
## [90] train-logloss:0.392891
## [91] train-logloss:0.392794
## [92] train-logloss:0.392579
## [93] train-logloss:0.392360
## [94] train-logloss:0.392196
## [95] train-logloss:0.392095
## [96] train-logloss:0.391964
## [97] train-logloss:0.391880
## [98] train-logloss:0.391759
## [99] train-logloss:0.391686
## [100] train-logloss:0.391538
```

```
# Predict on test data
preds <- predict(model_xgb, dtest)
# Extract true labels from the DMatrix
true_labels <- getinfo(dtest, "label")

# Compute the ROC curve
roc_obj_xgb <- roc(true_labels, preds)
```

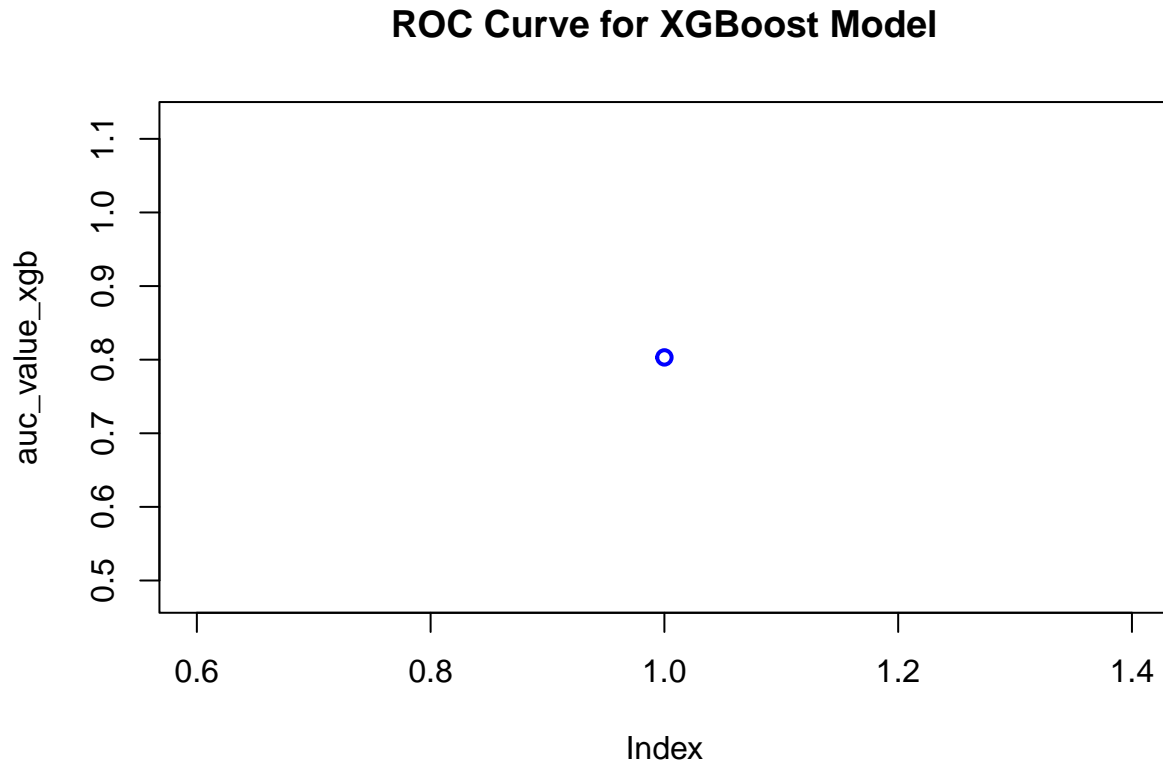
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
# Compute the AUC
auc_value_xgb <- as.numeric(auc(roc_obj_xgb))
print(auc_value_xgb)
```

```
## [1] 0.803075
```

```
# Plot the ROC curve
plot(auc_value_xgb, main = "ROC Curve for XGBoost Model", col = "blue", lwd = 2)
```



```
#####cross validation for XGBoost model#####
```

```
# Train the XGBoost model using caret
cv_model_xgb <- train(
  colon_scr_r ~ food_insecure_r + age + bmocat +
    female + racenew_r1 +
    poverty_r + insurance,
  data = nha3_colon_c,
  method = "xgbTree",          # Specify XGBoost as the method
  trControl = train_control,   # Use cross-validation setup
  metric = "ROC",              # Optimize based on ROC AUC
  tuneLength = 3               # Test 3 random parameter combinations (optional)
)
```

```
## [04:06:41] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:06:41] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:06:41] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:06:41] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:06:42] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:06:42] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:06:42] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:06:42] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]


```
## [04:13:46] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:46] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:46] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:46] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:48] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:48] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:48] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:48] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:49] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:49] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:49] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:49] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:51] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:51] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:51] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:52] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:52] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:52] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:52] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:54] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:54] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:54] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:54] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:55] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:55] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:55] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
## [04:13:55] WARNING: src/c_api/c_api.cc:935: 'ntree_limit' is deprecated, use 'iteration_range' instead
```

```
# Print the model summary
```

```
print(cv_model_xgb)
```

```
## eXtreme Gradient Boosting
```

```
##
```

```
## 48676 samples
```

```
## 7 predictor
```

```
## 2 classes: 'No', 'Yes'
```

```
##
```

```
## No pre-processing
```

```
## Resampling: Cross-Validated (10 fold)
```

```
## Summary of sample sizes: 43809, 43808, 43809, 43808, 43809, 43808, ...
```

```
## Resampling results across tuning parameters:
```

```
##
```

##	eta	max_depth	colsample_bytree	subsample	nrounds	ROC	Sens
##	0.3	1	0.6	0.50	50	0.8067484	0.5189019
##	0.3	1	0.6	0.50	100	0.8082346	0.5146834
##	0.3	1	0.6	0.50	150	0.8083299	0.5131647
##	0.3	1	0.6	0.75	50	0.8071119	0.5139242
##	0.3	1	0.6	0.75	100	0.8082923	0.5129112
##	0.3	1	0.6	0.75	150	0.8086110	0.5127429
##	0.3	1	0.6	1.00	50	0.8069735	0.5167921
##	0.3	1	0.6	1.00	100	0.8082859	0.5153578
##	0.3	1	0.6	1.00	150	0.8085385	0.5149361
##	0.3	1	0.8	0.50	50	0.8074304	0.5166229

##	0.3	1	0.8	0.50	100	0.8082768	0.5144301
##	0.3	1	0.8	0.50	150	0.8083759	0.5139239
##	0.3	1	0.8	0.75	50	0.8077018	0.5138385
##	0.3	1	0.8	0.75	100	0.8085343	0.5138392
##	0.3	1	0.8	0.75	150	0.8086751	0.5121513
##	0.3	1	0.8	1.00	50	0.8073419	0.5150200
##	0.3	1	0.8	1.00	100	0.8084560	0.5140077
##	0.3	1	0.8	1.00	150	0.8086671	0.5139233
##	0.3	2	0.6	0.50	50	0.8105805	0.5179732
##	0.3	2	0.6	0.50	100	0.8108874	0.5208421
##	0.3	2	0.6	0.50	150	0.8109067	0.5214334
##	0.3	2	0.6	0.75	50	0.8107483	0.5176359
##	0.3	2	0.6	0.75	100	0.8112520	0.5185638
##	0.3	2	0.6	0.75	150	0.8110545	0.5194927
##	0.3	2	0.6	1.00	50	0.8108538	0.5161171
##	0.3	2	0.6	1.00	100	0.8115886	0.5176366
##	0.3	2	0.6	1.00	150	0.8115290	0.5185644
##	0.3	2	0.8	0.50	50	0.8104560	0.5167081
##	0.3	2	0.8	0.50	100	0.8109526	0.5189018
##	0.3	2	0.8	0.50	150	0.8106990	0.5198299
##	0.3	2	0.8	0.75	50	0.8109128	0.5192390
##	0.3	2	0.8	0.75	100	0.8113094	0.5208430
##	0.3	2	0.8	0.75	150	0.8110562	0.5206735
##	0.3	2	0.8	1.00	50	0.8106001	0.5176364
##	0.3	2	0.8	1.00	100	0.8112911	0.5186490
##	0.3	2	0.8	1.00	150	0.8113601	0.5207584
##	0.3	3	0.6	0.50	50	0.8108029	0.5212646
##	0.3	3	0.6	0.50	100	0.8104045	0.5214330
##	0.3	3	0.6	0.50	150	0.8096866	0.5206738
##	0.3	3	0.6	0.75	50	0.8107995	0.5225304
##	0.3	3	0.6	0.75	100	0.8103844	0.5216864
##	0.3	3	0.6	0.75	150	0.8100603	0.5209268
##	0.3	3	0.6	1.00	50	0.8113923	0.5205057
##	0.3	3	0.6	1.00	100	0.8109846	0.5199148
##	0.3	3	0.6	1.00	150	0.8103803	0.5201678
##	0.3	3	0.8	0.50	50	0.8106817	0.5222773
##	0.3	3	0.8	0.50	100	0.8100270	0.5224462
##	0.3	3	0.8	0.50	150	0.8094262	0.5197461
##	0.3	3	0.8	0.75	50	0.8109685	0.5224454
##	0.3	3	0.8	0.75	100	0.8102883	0.5210116
##	0.3	3	0.8	0.75	150	0.8095474	0.5202515
##	0.3	3	0.8	1.00	50	0.8112891	0.5241330
##	0.3	3	0.8	1.00	100	0.8107203	0.5225302
##	0.3	3	0.8	1.00	150	0.8099810	0.5222775
##	0.4	1	0.6	0.50	50	0.8070431	0.5173835
##	0.4	1	0.6	0.50	100	0.8080150	0.5129960
##	0.4	1	0.6	0.50	150	0.8081997	0.5118141
##	0.4	1	0.6	0.75	50	0.8076711	0.5144297
##	0.4	1	0.6	0.75	100	0.8083614	0.5100425
##	0.4	1	0.6	0.75	150	0.8084790	0.5104640
##	0.4	1	0.6	1.00	50	0.8077775	0.5145133
##	0.4	1	0.6	1.00	100	0.8083600	0.5123200
##	0.4	1	0.6	1.00	150	0.8086406	0.5107172
##	0.4	1	0.8	0.50	50	0.8079248	0.5161176

##	0.4	1	0.8	0.50	100	0.8083660	0.5128276
##	0.4	1	0.8	0.50	150	0.8086167	0.5124891
##	0.4	1	0.8	0.75	50	0.8079187	0.5183111
##	0.4	1	0.8	0.75	100	0.8085995	0.5110550
##	0.4	1	0.8	0.75	150	0.8087105	0.5115608
##	0.4	1	0.8	1.00	50	0.8076806	0.5122356
##	0.4	1	0.8	1.00	100	0.8083320	0.5135018
##	0.4	1	0.8	1.00	150	0.8085596	0.5100421
##	0.4	2	0.6	0.50	50	0.8100651	0.5185645
##	0.4	2	0.6	0.50	100	0.8104281	0.5187325
##	0.4	2	0.6	0.50	150	0.8105374	0.5204204
##	0.4	2	0.6	0.75	50	0.8105794	0.5183956
##	0.4	2	0.6	0.75	100	0.8109815	0.5208428
##	0.4	2	0.6	0.75	150	0.8108664	0.5197460
##	0.4	2	0.6	1.00	50	0.8107463	0.5182264
##	0.4	2	0.6	1.00	100	0.8111953	0.5216864
##	0.4	2	0.6	1.00	150	0.8111485	0.5215176
##	0.4	2	0.8	0.50	50	0.8105842	0.5183970
##	0.4	2	0.8	0.50	100	0.8106420	0.5216873
##	0.4	2	0.8	0.50	150	0.8103925	0.5205061
##	0.4	2	0.8	0.75	50	0.8106110	0.5199986
##	0.4	2	0.8	0.75	100	0.8112046	0.5188174
##	0.4	2	0.8	0.75	150	0.8108458	0.5201679
##	0.4	2	0.8	1.00	50	0.8111247	0.5195770
##	0.4	2	0.8	1.00	100	0.8113090	0.5201676
##	0.4	2	0.8	1.00	150	0.8112820	0.5205897
##	0.4	3	0.6	0.50	50	0.8104739	0.5214323
##	0.4	3	0.6	0.50	100	0.8091853	0.5226987
##	0.4	3	0.6	0.50	150	0.8085576	0.5192397
##	0.4	3	0.6	0.75	50	0.8102321	0.5212647
##	0.4	3	0.6	0.75	100	0.8097400	0.5203359
##	0.4	3	0.6	0.75	150	0.8089232	0.5219401
##	0.4	3	0.6	1.00	50	0.8107434	0.5224464
##	0.4	3	0.6	1.00	100	0.8100468	0.5219397
##	0.4	3	0.6	1.00	150	0.8094581	0.5216863
##	0.4	3	0.8	0.50	50	0.8093831	0.5223616
##	0.4	3	0.8	0.50	100	0.8086966	0.5216868
##	0.4	3	0.8	0.50	150	0.8077121	0.5184802
##	0.4	3	0.8	0.75	50	0.8103535	0.5204218
##	0.4	3	0.8	0.75	100	0.8096648	0.5208427
##	0.4	3	0.8	0.75	150	0.8085908	0.5218557
##	0.4	3	0.8	1.00	50	0.8109413	0.5228673
##	0.4	3	0.8	1.00	100	0.8100025	0.5235432
##	0.4	3	0.8	1.00	150	0.8092030	0.5226149
##	Spec						
##	0.9196994						
##	0.9210843						
##	0.9218176						
##	0.9208399						
##	0.9211387						
##	0.9214646						
##	0.9206770						
##	0.9208400						
##	0.9215189						

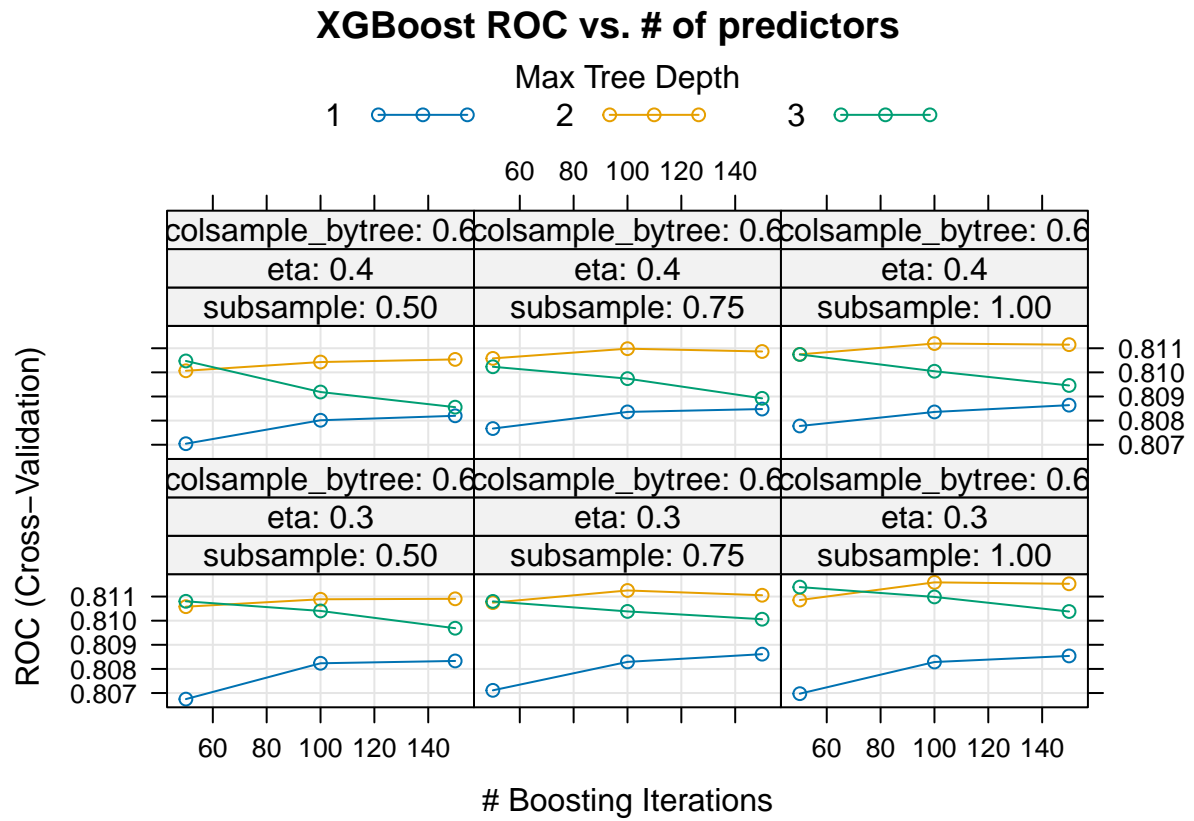
0.9203239
0.9209486
0.9216547
0.9207585
0.9216003
0.9222521
0.9208942
0.9211387
0.9216818
0.9208672
0.9209758
0.9208673
0.9213561
0.9214917
0.9208943
0.9212474
0.9213831
0.9212745
0.9207314
0.9206227
0.9206227
0.9205956
0.9202154
0.9203512
0.9209758
0.9214918
0.9205956
0.9202969
0.9195093
0.9199438
0.9199167
0.9199981
0.9197538
0.9204599
0.9201069
0.9202698
0.9202155
0.9197809
0.9196995
0.9195637
0.9196723
0.9196995
0.9192379
0.9194823
0.9196724
0.9198080
0.9211115
0.9216818
0.9203512
0.9217904
0.9221435
0.9203240
0.9210029
0.9217905

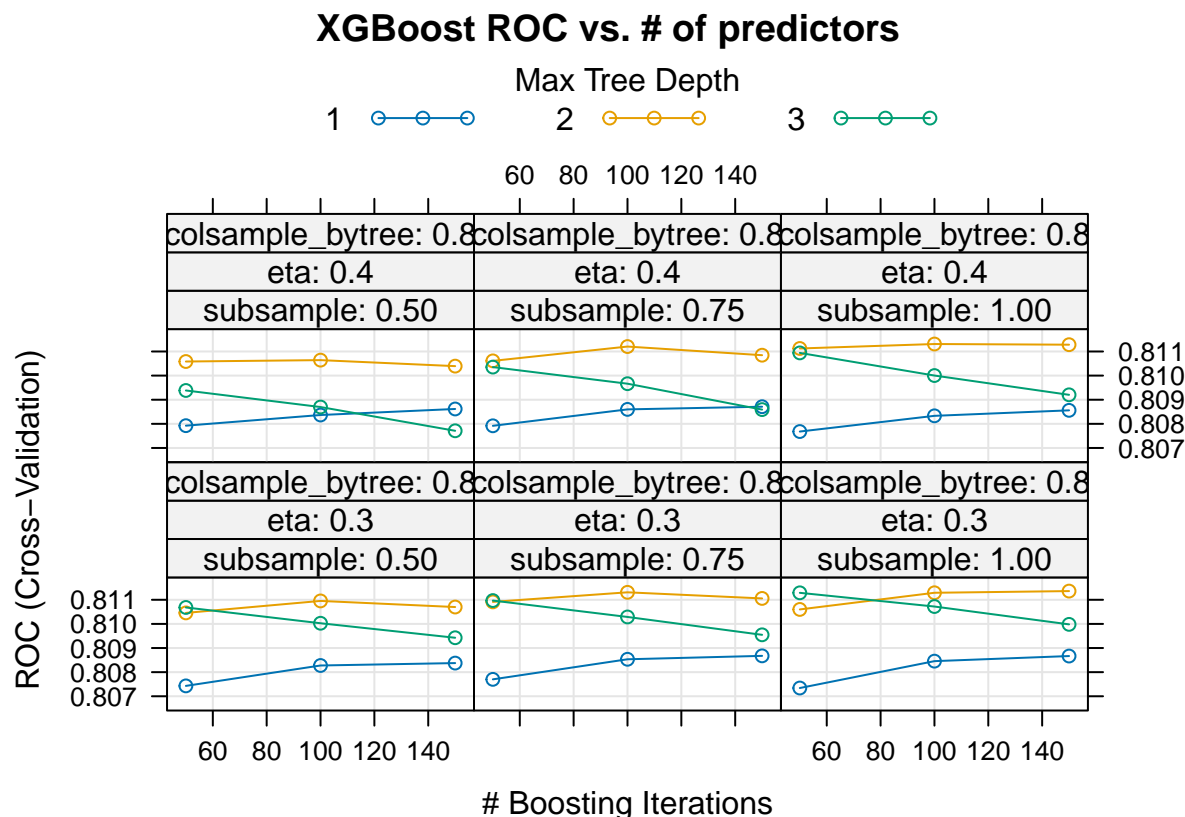

```

## 0.9204599
## 0.9216275
## 0.9222792
## 0.9193193
## 0.9213831
## 0.9218176
## 0.9209486
## 0.9210573
## 0.9220620
## 0.9209488
## 0.9211931
## 0.9210302
## 0.9210572
## 0.9204328
## 0.9207043
## 0.9211115
## 0.9207857
## 0.9202969
## 0.9208672
## 0.9206227
## 0.9201882
## 0.9202969
## 0.9208943
## 0.9207314
## 0.9207586
## 0.9207314
## 0.9201611
## 0.9199168
## 0.9194007
## 0.9201882
## 0.9200254
## 0.9199167
## 0.9200796
## 0.9197810
## 0.9198895
## 0.9193737
## 0.9199710
## 0.9192649
## 0.9198081
## 0.9200527
## 0.9199982
## 0.9192107
## 0.9195365
## 0.9191020
## 0.9194822
##
## Tuning parameter 'gamma' was held constant at a value of 0
## Tuning
## parameter 'min_child_weight' was held constant at a value of 1
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were nrounds = 100, max_depth = 2, eta
## = 0.3, gamma = 0, colsample_bytree = 0.6, min_child_weight = 1 and subsample
## = 1.

```

```
plot(cv_model_xgb, main = "XGBoost ROC vs. # of predictors")
```





```
extract_auc_xgb=as.data.frame(cv_model_xgb$resample)

auc_xgb_xvalidated<-mean(as.numeric(extract_auc_xgb$ROC))

#adding AUC metric to the dataframe
auc_table <- auc_table %>% add_row(Model="XGBoost model", AUC_single_model=auc_value_xgb,
                                   AUC_10F_xvalidated = auc_xgb_xvalidated)

auc_table
```

##	Model	AUC_single_model	AUC_10F_xvalidated
## 1	Logistic Regression model	0.7918455	0.7915490
## 2	Random Forest Model	0.8064965	0.7736533
## 3	Support Vector Machine model	0.7880106	0.8033707
## 4	XGBoost model	0.8030750	0.8115886

The Gradient Boosting Model performed better than any other model we tried with the single model AU-CROC 0.80 and 10-fold cross validated AUROC of 0.81.

Conclusion We developed predictive models for colon cancer screening using relevant social determinants of health and demographic characteristics, including food insecurity, poverty, health insurance, age, sex, and BMI. Four different models of varying complexity were tested: Logistic Regression, Random Forest, Support Vector Machine, and Gradient Boosting. Model performance was evaluated using the Area Under the Receiver Operating Characteristics Curve (AUROC). The 'Caret' package was employed to perform 10-fold cross-validation for each predictive model. Among the models, Gradient Boosting demonstrated the best performance, achieving a 10-fold cross-validated AUROC of 0.81.

Characteristic	N	Overall N = 176,355 ^I	1 N = 154,113 ^I	2 N = 9,631 ^I	3 N =
food_insecure	176,355	22,242 (13%)	0 (0%)	9,631 (100%)	12,611
Survey year	176,355				
2018		56,143 (32%)	48,055 (31%)	3,540 (37%)	4,548
2019		31,860 (18%)	27,755 (18%)	1,598 (17%)	2,507
2020		31,450 (18%)	27,998 (18%)	1,505 (16%)	1,947
2021		29,343 (17%)	26,365 (17%)	1,371 (14%)	1,607
2022		27,559 (16%)	23,940 (16%)	1,617 (17%)	2,002
Age	176,355	52 (36, 66)	53 (37, 67)	47 (32, 62)	47 (36, 66)
bmicat	141,986				
1		2,263 (1.6%)	1,944 (1.6%)	132 (1.8%)	187 (1.6%)
2		45,179 (32%)	40,711 (33%)	1,984 (27%)	2,484 (32%)
3		49,024 (35%)	43,777 (35%)	2,287 (31%)	2,960 (35%)
4		45,520 (32%)	38,046 (31%)	3,071 (41%)	4,403 (32%)
female	176,346	94,457 (54%)	81,342 (53%)	5,644 (59%)	7,471 (54%)
racenew_r1	170,436				
Asian		10,405 (6.1%)	9,479 (6.3%)	488 (5.4%)	438 (6.1%)
Black		19,658 (12%)	14,978 (10%)	1,811 (20%)	2,869 (12%)
Other		5,693 (3.3%)	4,329 (2.9%)	500 (5.5%)	864 (3.3%)
White		134,680 (79%)	120,650 (81%)	6,264 (69%)	7,766 (79%)
white	176,355	134,680 (76%)	120,650 (78%)	6,264 (65%)	7,766 (76%)
black	176,355	19,658 (11%)	14,978 (9.7%)	1,811 (19%)	2,869 (11%)
asian	176,355	10,405 (5.9%)	9,479 (6.2%)	488 (5.1%)	438 (5.9%)
other_race	176,355	5,693 (3.2%)	4,329 (2.8%)	500 (5.2%)	864 (3.2%)
hispanic	176,139	24,163 (14%)	19,232 (12%)	2,250 (23%)	2,681 (14%)
foodstamp	171,215	18,821 (11%)	10,568 (7.1%)	3,031 (32%)	5,222 (11%)
food_security_score	176,355	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	1.00 (1.00, 2.00)	5.00 (3.00, 7.00)
private_ins	175,617	103,752 (59%)	96,542 (63%)	3,717 (39%)	3,493 (59%)
state_ins	175,319	4,878 (2.8%)	3,664 (2.4%)	493 (5.2%)	721 (2.8%)
chi_ins	175,319	101 (<0.1%)	78 (<0.1%)	10 (0.1%)	13 (<0.1%)
medicaid	175,319	17,471 (10.0%)	11,259 (7.3%)	2,295 (24%)	3,917 (10.0%)
medicare	175,319	49,533 (28%)	43,925 (29%)	2,364 (25%)	3,244 (28%)
other_gov_ins	175,319	2,221 (1.3%)	1,783 (1.2%)	191 (2.0%)	247 (1.3%)
no_ins	175,319	14,491 (8.3%)	10,644 (6.9%)	1,488 (16%)	2,359 (8.3%)
insurance	175,617				
1		103,752 (59%)	96,542 (63%)	3,717 (39%)	3,493 (59%)
2		52,663 (30%)	42,148 (27%)	4,123 (43%)	6,392 (30%)
3		19,202 (11%)	14,827 (9.7%)	1,727 (18%)	2,648 (11%)
poverty_r	172,409				
<1		16,981 (9.8%)	10,342 (6.9%)	2,312 (25%)	4,327 (9.8%)
>=5		54,012 (31%)	53,086 (35%)	599 (6.4%)	327 (31%)
1-1.99		29,634 (17%)	21,807 (14%)	3,149 (34%)	4,678 (17%)
2-4.99		71,782 (42%)	65,479 (43%)	3,289 (35%)	3,014 (42%)
health_r1	176,237				
Excellent		41,936 (24%)	39,039 (25%)	1,492 (16%)	1,405 (24%)
Fair		19,352 (11%)	14,265 (9.3%)	1,852 (19%)	3,235 (11%)
Good		49,460 (28%)	42,311 (27%)	3,175 (33%)	3,994 (28%)
Poor		6,022 (3.4%)	3,901 (2.5%)	594 (6.2%)	1,527 (3.4%)
Very Good		59,447 (34%)	54,504 (35%)	2,510 (26%)	2,433 (34%)
colon_scr	51,957	39,059 (75%)	35,182 (76%)	1,594 (68%)	2,283 (75%)