Supervised machine learning techniques to predict depression in ambulatory care setting

Sharmin Shabnam

1/10/2020

Contents

1 Introduction

Depression is a common mental health disorder affecting more than 300 million people worldwide.[1] It is one of the leading causes of disability and premature mortality preventing people from reaching their full potential.[1,2] One in every twelve American adults aged 20 and above had depression in a given two-week period during 2013-2016.[3] In 2017, more than 17 million adults in the United States were estimated to have at least one major depressive episode.[4]

Depression disproportionately affects vulnerable subgroups of the population. For example, women are twice as likely as men to have depression.[3] Lower socioeconomic status was also found to be associated with higher rate of depression.[5] Despite the availability of successful psychological and pharmacological treatments for moderate and severe depression,[6] an estimated 56.3% of people with depression left untreated.[7]

With a view to promoting remission, preventing relapse, and reducing emotional and financial burden of mental health diseases, it is imperative to emphasize on early detection, intervention, and appropriate treatment of depression.[8] An increasingly available large electronic medical records made it possible to apply advanced analytic approach to predict a range of health conditions.[9] In some cases, machine learning techniques may be able to outperform conventional discourse of clinical diagnosis and prognosis.[10,11] The aim of this study was to predict depression with the available data in the US ambulatory healthcare setting with the application of a range of machine learning techniques.

2 Data Exploration

Data Source: The data for this study was taken from the US National Ambulatory Medical Care Survey (NAMCS). The survey collects data on a national sample of ambulatory care services in the emergency and outpatient departments, and ambulatory surgery locations of noninstitutional general and short-stay hospitals. The details of the survey have been documented elsewhere.[12]

For the purpose of this study, the 2014 cycle of the survey was used because this was the first (and the largest) NAMCS survey with all the required variables, especially the comorbidities and risk behaviours postulated to be associated with depression.

Outcome variable: The outcome variable, depression, was based on the survey question whether, regardless of other diagnoses elsewhere recorded, the participants currently had depression at that time. In this survey, depression "includes affective disorders and major depressive disorders, such as episodes of depressive reaction, psychogenic depression, and reactive depression."[12]

Predictors (features): To make the prediction models clinically relevant, the predictors were selected on previous literature.[1-8] The predictors considered in the models include:

- Demographic variables:
  - Patient's age: Since age was available as both numeric and categorical, an exploratory logistic regression analysis was conducted with depression as the outcome variable. The model with age as a categorical variable had lower AIC (Akaike Information Criteria), and so it was later used in all the analysis.
  - Sex: Male and female.
  - Race: White, Black, and Other.
  - Body-mass index (BMI): Classified as Underweight (BMI <18.5), normal weight (BMI: 18.5-24.9), overweight and obesity (BMI: 25 and above), and missing or unknown. Insurance type: Private insurance, Medicare, Medicaid, Other, and Unknown or missing Geographic region in USA: Northeast, Midwest, South, and West
- Risk behaviours
  - Tobacco use: Never, Former, Current, Unknown or missing.
  - Substance abuse: Yes/no.
  - Alcohol misuse, abuse or dependence: Yes/no.
  - History of medication use: Number of medications used.
- Comorbidities (whether the patient currently had):
  - Alzheimer's disease
  - Arthritis
  - Asthma
  - Depression
  - Cancer
  - Cerebrovascular disease
  - Chronic kidney disease
  - Chronic obstructive pulmonary disease
  - Congestive heart failure
  - Coronary artery disease
  - Diabetes mellitus Type 1
  - Diabetes mellitus Type 2
  - End-stage renal disease
  - Pulmonary embolism or deep vein thrombosis
  - HIV infection
  - Hyperlipidemia

- Hypertension
- Obstructive sleep apnea
- Osteoporosis

The descriptive statistics of these variables are presented in Table 1.

```
#Load data file from github repository
namcs2014 <- read.csv("https://raw.githubusercontent.com/shabnam-shbd/
Data_Science_Capstone_Project_CYO/master/namcs2014-stata.csv
", header=TRUE, sep=",")


#Select variables from the data set and sort according
#to their types (factors and numeric variables)
selected_vars <- c("AGE", "AGER", "SEX", "RACER","PAYTYPER", "BMI", "USETOBAC",
                   "SUBSTAB", "ETOHAB","ALZHD", "ARTHRTIS","ASTHMA","DEPRN",
                   "CANCER", "CEBVD", "CKD","COPD","CHF","CAD","DIABTYP1",
                   "DIABTYP2","ESRD","HPE","HIV", "HYPLIPID","HTN","OSA",
                   "OSTPRSIS", "NUMMED", "REGIONOFF")

#Create dataframe for descriptive table
namcs_table1 <- namcs2014 %>%
 select(selected_vars) %>%
 #Convert BMI to numeric variable
 mutate(BMI = as.numeric(BMI)) %>%
 #Transform BMI from numeric to categorical variable
 mutate(BMI=cut(BMI, breaks=c(-10,0,18.5,25,81.2),
                labels=c("Unknown or Missing","Underweight",
                         "Normal weight","Overweight or Obese"),
                right=FALSE)) %>%
  #Transform PAYTYPER into a categorical variable
  mutate(PAYTYPER=cut(PAYTYPER, breaks=c(-10,1,2,3,4,8),
                      labels=c("Unknown or missing",
                               "Private insurance",
                               "Medicare","Medicaid",
                               "Other"),
                      right=FALSE)) %>%
  #Transform USETOBAC into a categorical variable
  mutate(USETOBAC=cut(USETOBAC, breaks=c(-10,1,2,3,4),
                      labels=c("Unknown or missing",
                               "Never","Former","Current"),
                      right=FALSE))
#Sort variables according to their types (factor,
#categorical and numeric variables)
factor_vars <- c("AGER","SEX", "RACER","PAYTYPER", "BMI", "USETOBAC",
                 "SUBSTAB", "ETOHAB","ALZHD","ARTHRTIS","ASTHMA",
                 "DEPRN","CANCER", "CEBVD", "CKD","COPD","CHF","CAD",
                 "DIABTYP1","DIABTYP2","ESRD","HPE","HIV", "HYPLIPID",
                 "HTN","OSA","OSTPRSIS", "REGIONOFF")

namcs_table1[factor_vars] <- lapply(namcs_table1[factor_vars], as.factor)

yesno_vars <- c("ETOHAB","ALZHD","ARTHRTIS","ASTHMA","DEPRN","CANCER",
                "CEBVD", "CKD","COPD","CHF","CAD","DIABTYP1","DIABTYP2",
                "ESRD","HPE","HIV","HYPLIPID","HTN","OSA","OSTPRSIS","SUBSTAB")
```

```r
namcs_table1[yesno_vars] <- lapply(namcs_table1[yesno_vars],factor,
                                   levels = c(1, 0),
                                   labels = c("Yes", "No"))

numeric_vars <- c("AGE","NUMMED")

namcs_table1[numeric_vars] <- sapply(namcs_table1[numeric_vars],as.numeric)

levels(namcs_table1$SEX) <- list("Female"="1", "Male"="2")

levels(namcs_table1$REGIONOFF) <- list("Northeast"="1", "Midwest"="2",
                                       "South"="3", "West"="4")

levels(namcs_table1$AGER) <- list("<15 years"="1", "15-24 years"="2",
                                  "25-44 years"="3","45-64 years"="4",
                                  "65-74 years"="5","75 years and above"="6")

levels(namcs_table1$RACER) <- list("White"="1", "Black"="2", "Other"="3")

var_labels <- c(AGE = "Patient age (years)",
                AGER = "Patient age categories",
                SEX = "Sex",
                RACER = "Race",
                PAYTYPER = "Insurance type",
                BMI = "Body-mass index category",
                USETOBAC = "Tobacco use",
                SUBSTAB = "Substance abuse",
                ETOHAB = "Alcohol misuse, abuse or dependence",
                ALZHD = "Alzheimer\'s disease",
                ARTHRTIS = "Arthritis",
                ASTHMA = "Asthma",
                DEPRN = "Depression",
                CANCER = "Cancer",
                CEBVD = "Cerebrovascular disease",
                CKD = "Chronic kidney disease",
                COPD = "Chronic obstructive pulmonary disease",
                CHF = "Congestive heart failure",
                CAD = "Coronary artery disease",
                DIABTYP1 = "Diabetes mellitus Type 1",
                DIABTYP2 = "Diabetes mellitus Type 2",
                ESRD = "End-stage renal disease",
                HPE = "Pulmonary embolism or deep vein thrombosis",
                HIV = "HIV infection",
                HYPLIPID = "Hyperlipidemia",
                HTN = "Hypertension",
                OSA = "Obstructive sleep apnea",
                OSTPRSIS = "Osteoporosis",
                NUMMED = "Number of medications",
                REGIONOFF = "Geographic region in USA")

#Assign variable labels to each variable
label(namcs_table1) = lapply(names(namcs_table1),
                             function(x) var_labels[match(x, names(var_labels))])
```

Table 1: Descriptive Statistics of Selected Variables in NAMCS 2014 Data Set.

```
# To print nicely formatted table of descriptive statistics
print(dfSummary(namcs_table1, plain.ascii = F, style = "grid",
                subtitle.emphasis = T, varnumbers = F, labels.col = T,
                graph.col = F, headings = F, display.labels = F,
                valid.col = F, na.col = F, tmp.img.dir = "/tmp"))
```

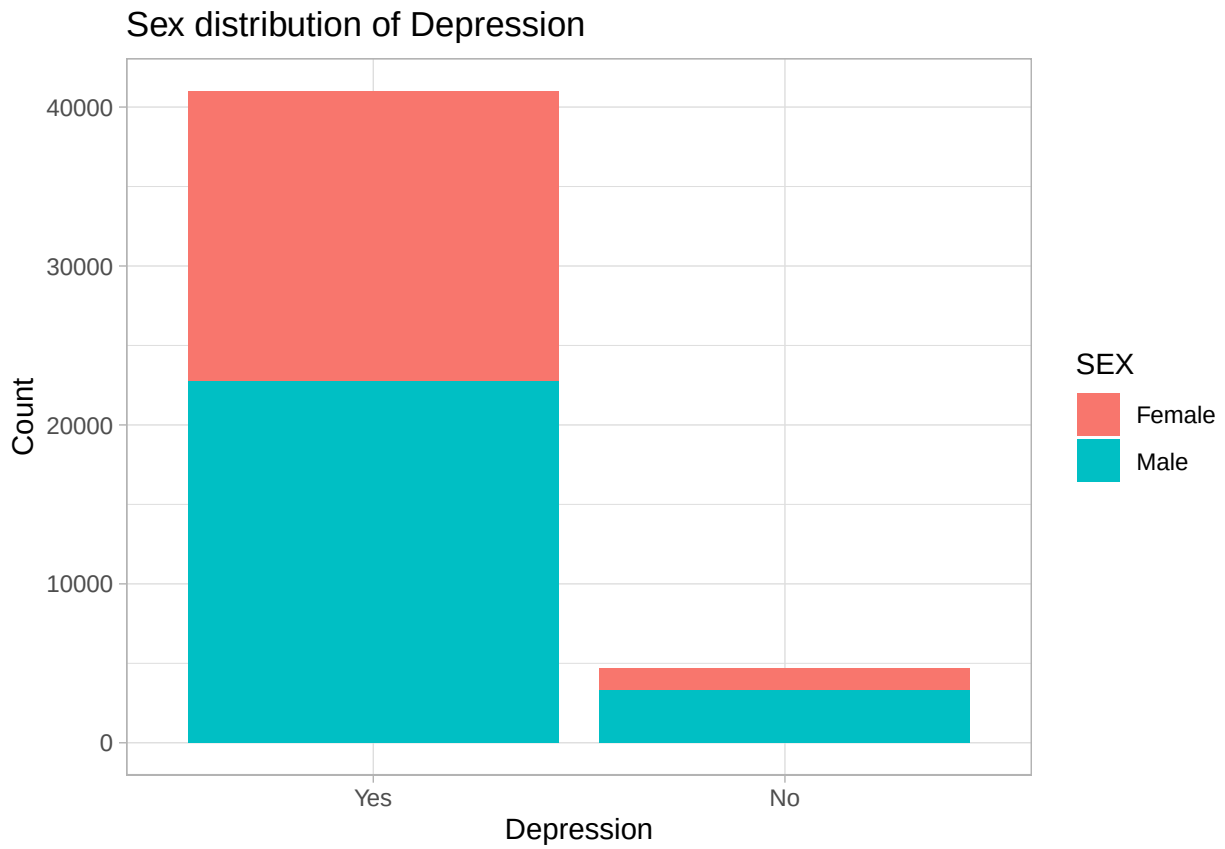
Variable	Label	Stats / Values	Freqs (% of Valid)
AGE [labelled, numeric]	Patient age (years)	Mean (sd) : 47.4 (24.9) min < med < max: 0 < 52 < 92 IQR (CV) : 38 (0.5)	93 distinct values
AGER [labelled, factor]	Patient age categories	1. <15 years 2. 15-24 years 3. 25-44 years 4. 45-64 years 5. 65-74 years 6. 75 years and above	6680 (14.6%) 3281 (7.2%) 8484 (18.6%) 13787 (30.2%) 7178 (15.7%) 6300 (13.8%)
SEX [labelled, factor]	Sex	1. Female 2. Male	26137 (57.2%) 19573 (42.8%)
RACER [labelled, factor]	Race	1. White 2. Black 3. Other	39371 (86.1%) 4138 (9.0%) 2201 (4.8%)
PAYTYPER [labelled, factor]	Insurance type	1. Unknown or missing 2. Private insurance 3. Medicare 4. Medicaid 5. Other	3104 (6.8%) 22422 (49.0%) 12076 (26.4%) 5042 (11.0%) 3066 (6.7%)
BMI [labelled, factor]	Body-mass index category	1. Unknown or Missing 2. Underweight 3. Normal weight 4. Overweight or Obese	18128 (39.7%) 2123 (4.6%) 7788 (17.0%) 17671 (38.7%)
USETOBAC [labelled, factor]	Tobacco use	1. Unknown or missing 2. Never 3. Former 4. Current	9768 (21.4%) 24512 (53.6%) 6990 (15.3%) 4440 (9.7%)
SUBSTAB [labelled, factor]	Substance abuse	1. Yes 2. No	1093 (2.4%) 44617 (97.6%)
ETOHAB [labelled, factor]	Alcohol misuse, abuse or dependence	1. Yes 2. No	413 (0.9%) 45297 (99.1%)
ALZHD [labelled, factor]	Alzheimer's disease	1. Yes 2. No	341 (0.8%) 45369 (99.2%)
ARTHRTIS [labelled, factor]	Arthritis	1. Yes 2. No	6462 (14.1%) 39248 (85.9%)
ASTHMA [labelled, factor]	Asthma	1. Yes 2. No	3062 (6.7%) 42648 (93.3%)
DEPRN [labelled, factor]	Depression	1. Yes 2. No	4687 (10.2%) 41023 (89.8%)
CANCER [labelled, factor]	Cancer	1. Yes 2. No	3438 (7.5%) 42272 (92.5%)
CEBVD [labelled, factor]	Cerebrovascular disease	1. Yes 2. No	906 (2.0%) 44804 (98.0%)

Variable	Label	Stats / Values	Freqs (% of Valid)
CKD	Chronic kidney disease	1. Yes	1280 (2.8%)
[labelled, factor]		2. No	44430 (97.2%)
COPD	Chronic obstructive pulmonary	1. Yes	1828 (4.0%)
[labelled, factor]	disease	2. No	43882 (96.0%)
CHF	Congestive heart failure	1. Yes	769 (1.7%)
[labelled, factor]		2. No	44941 (98.3%)
CAD	Coronary artery disease	1. Yes	3024 (6.6%)
[labelled, factor]		2. No	42686 (93.4%)
DIABTYP1	Diabetes mellitus Type 1	1. Yes	306 (0.7%)
[labelled, factor]		2. No	45404 (99.3%)
DIABTYP2	Diabetes mellitus Type 2	1. Yes	3214 (7.0%)
[labelled, factor]		2. No	42496 (93.0%)
ESRD	End-stage renal disease	1. Yes	131 (0.3%)
[labelled, factor]		2. No	45579 (99.7%)
HPE	Pulmonary embolism or deep vein	1. Yes	280 (0.6%)
[labelled, factor]	thrombosis	2. No	45430 (99.4%)
HIV	HIV infection	1. Yes	147 (0.3%)
[labelled, factor]		2. No	45563 (99.7%)
HYPLIPID	Hyperlipidemia	1. Yes	8574 (18.8%)
[labelled, factor]		2. No	37136 (81.2%)
HTN	Hypertension	1. Yes	13217 (28.9%)
[labelled, factor]		2. No	32493 (71.1%)
OSA	Obstructive sleep apnea	1. Yes	1340 (2.9%)
[labelled, factor]		2. No	44370 (97.1%)
OSTPRSIS	Osteoporosis	1. Yes	1064 (2.3%)
[labelled, factor]		2. No	44646 (97.7%)
NUMMED	Number of medications	Mean (sd) : 3.8 (4.5)	31 distinct values
[labelled, numeric]		min < med < max:	
		0 < 2 < 30	
		IQR (CV) : 5 (1.2)	
REGIONOFF	Geographic region in USA	1. Northeast	6571 (14.4%)
[labelled, factor]		2. Midwest	12622 (27.6%)
		3. South	16012 (35.0%)
		4. West	10505 (23.0%)

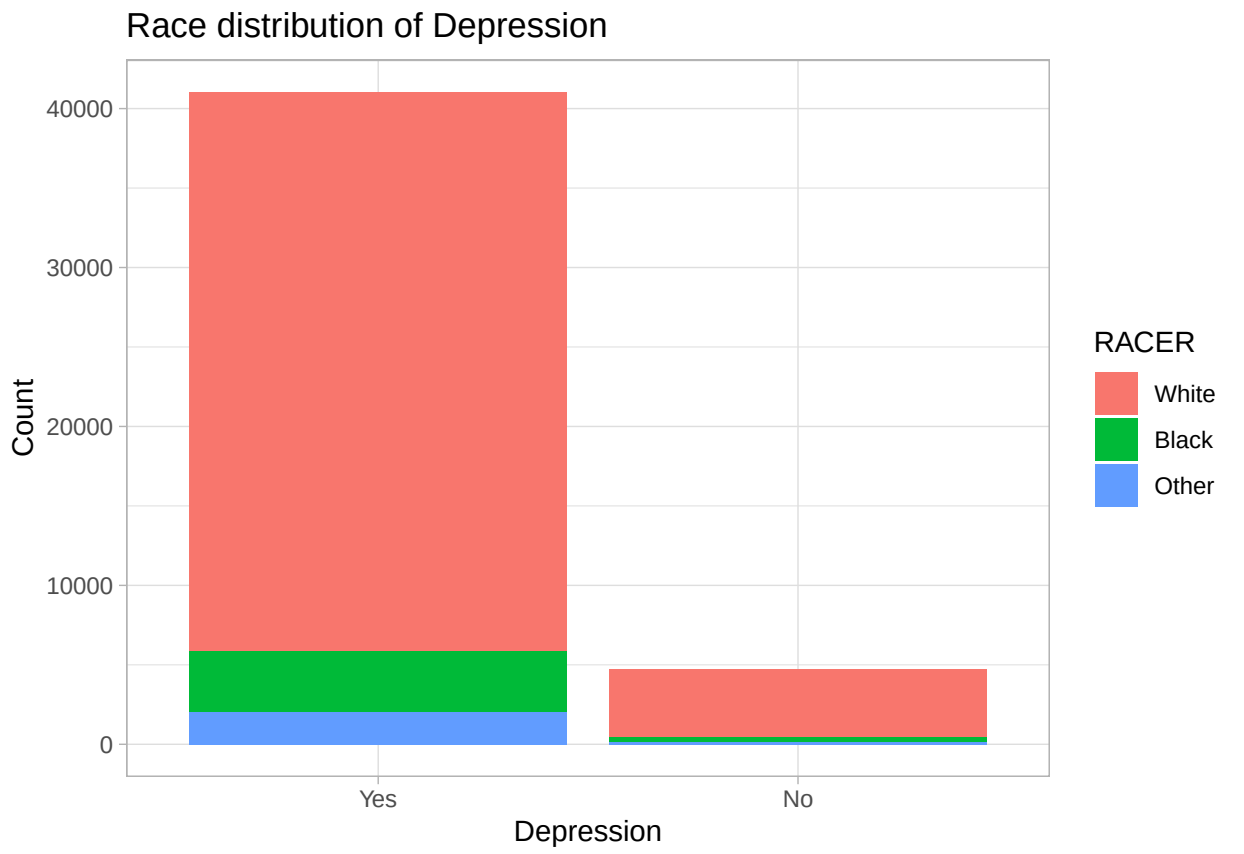
The following graphs show sex and race distribution of the depression status.

```
#Select and transform categorical variables from the data set
namcs_clean <- namcs2014 %>%
  mutate(BMI = as.numeric(BMI)) %>%
  select(selected_vars) %>%
  mutate(SEX = factor(ifelse(SEX == 1,1,0))) %>%
  mutate(BMI=cut(BMI, breaks=c(-10,0,18.5,25,81.2),
                 labels=c("1","2","3","4"), right=FALSE)) %>%
  mutate(PAYTYPER=cut(PAYTYPER, breaks=c(-10,1,2,3,4,8),
                      labels=c("1", "2", "3","4","5"), right=FALSE)) %>%
  mutate(USETOBAC=cut(USETOBAC, breaks=c(-10,1,2,3,4),
                      labels=c("1", "2","3","4"), right=FALSE))

#Sex distribution of Depression
namcs_clean %>%
  mutate(SEX = factor(ifelse(SEX == 1,1,0), labels = c("Female", "Male"))) %>%
```

```
  ggplot(aes(x = as.factor(DEPRN), fill= SEX)) + theme_light() +
  geom_bar() +
  labs(title="Sex distribution of Depression", x="Depression", y="Count") +
  scale_x_discrete( breaks = c("0","1"),labels= c("Yes","No"))
```




```
#Race distribution of Depression
namcs_clean %>%
  mutate(RACER=cut(as.numeric(RACER), breaks=c(1,2,3,4),
                   labels=c("White", "Black","Other"), right=FALSE)) %>%
  ggplot(aes(x = as.factor(DEPRN), fill= RACER)) + theme_light() +
  geom_bar() +
  labs(title="Race distribution of Depression", x="Depression", y="Count") +
  scale_x_discrete( breaks = c("0","1"),labels= c("Yes","No"))
```



2.1 Data Prepration for Final Analysis

The data was thoroughly checked for data formatting, missingness and outliers. All but one (number of medications used) of the predictors was a continuous variable. All the comorbidities were coded as binary (yes/no) variables, and did not have any missing data.

Insurance type was further reclassified as stated above. Worker's compensation, self-pay, no charge/charity, and other were grouped into "other" category (as used in previous literature).[13] Missing and unknown categories were merged into a separate category.[13] The latter methodology was applied also to BMI, and tobacco use.

```
#Factorize some variables
factor_vars <- c( 'REGIONOFF',"AGER", 'RACER','PAYTYPER',"USETOBAC",'BMI')
namcs_clean[factor_vars] <- lapply(namcs_clean[factor_vars], as.factor)

yesno_vars <- c("ETOHAB","ALZHD","ARTHRTIS","ASTHMA","DEPRN","CANCER",
                "CEBVD", "CKD","COPD","CHF","CAD","DIABTYP1","DIABTYP2",
                "ESRD","HPE","HIV","HYPLIPID","HTN","OSA","OSTPRSIS","SUBSTAB")

namcs_clean[yesno_vars] <- lapply(namcs_clean[yesno_vars],factor,
                                  levels = c(0,1))

#Select reference catergory
namcs_clean$BMI <- relevel(namcs_clean$BMI, ref = "3")
namcs_clean$PAYTYPER <- relevel(namcs_clean$PAYTYPER, ref = "2")
namcs_clean$USETOBAC <- relevel(namcs_clean$USETOBAC, ref = "2")
```

```
#Clean up memory
rm(namcs2014, namcs_table1,factor_vars,selected_vars)
invisible(gc())
```

2.2 Creating partitions of training and test data set

The data set was divided into two parts: train (80%) and test (20%) dataset.

```
#Create Data partition into test set and training set 80% and 20%
set.seed(1000, sample.kind="Rounding") #if using R 3.6 or later
test_index <- createDataPartition(namcs_clean$DEPRN,
                                  times = 1, p = 0.2, list = FALSE)

#Test and Training set for final analysis
test_set <- namcs_clean[test_index, ] %>%  select(-AGE)
train_set <- namcs_clean[-test_index, ] %>% select(-AGE)
dim(test_set)
```

[1] 9143 29

```
dim(train_set)
```

[1] 36567 29

```
table(test_set$DEPRN)
```

0 1 8205 938

```
table(train_set$DEPRN)
```

```
0     1
```

32818 3749

```
#Clean up memory
rm(namcs,factor_vars,selected_vars)
invisible(gc())
```

3 Methods and Analysis

This project compared the predicting abilities of different models and their performances were compared in terms of area under the receiver operating characteristics (ROC) curve (AUC), sensitivity, specificity, and overall accuracy measures. The following algorithms were compared:

[a] Multivariable logistic regression using all the predictors listed above. The best threshold for outcome classification was estimated based on Youden method which is widely used in the literature.[14] Based on the threshold, the AUC statistics and other measures of interests were reported.

[b] Classification and Regression Tree (CART) with a 5-fold cross-validation with a minimum split size of 20 and complexity parameter (cp) of 0.00001.

[c] Random Forest with 500 trees where 5 variables were randomly selected to develop each tree.

3.1 Logistic Regression Model

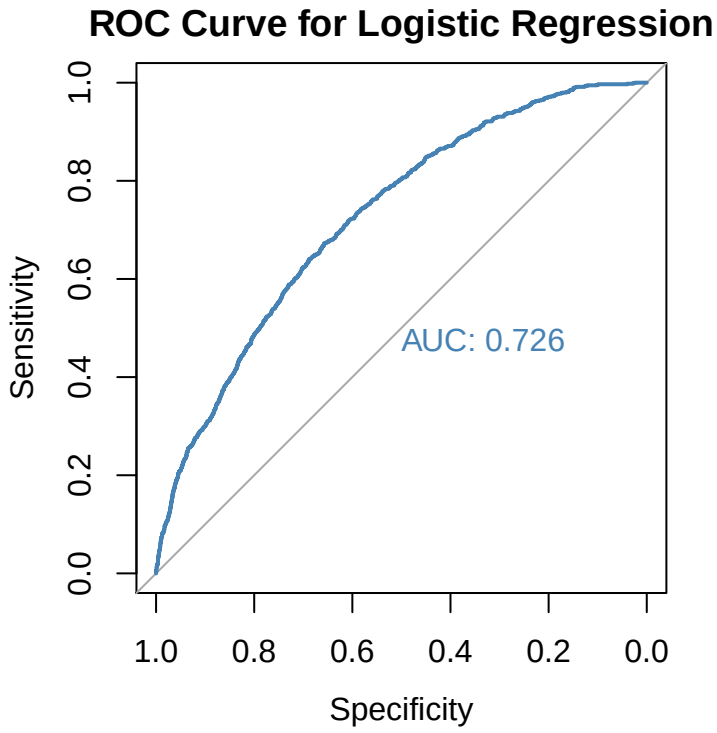
The first model used in this project is Logistic Reression which is one of the most commonly used methods for predictive algorithms. The following code implements the *glm* (Generalized Linear Models) function in *R*.

```r
#Training a logistic regression model with the caret glm method
set.seed(1000, sample.kind = "Rounding")
fit_logistic_reg <- glm(DEPRN ~ .,
                        family=binomial(link="logit"),
                        data=train_set)

#ROC graph and area under the curve
pred_logistic <- predict(fit_logistic_reg,
                         newdata = test_set, type = "response")
roc_lr <- roc(response = test_set$DEPRN, predictor = pred_logistic)
```

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

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

```r
plot(roc_lr, print.auc=TRUE,
     lwd=2, xlim=c(1,0),
     col="steelblue",
     main="ROC Curve for Logistic Regression")
```




```r
auc_lr <- auc(roc_lr)
auc_lr
```

Area under the curve: 0.7258

```r
# Threshold using youden method
threshold_youden_lr <- coords(roc_lr, "best",
                              ret = c("threshold", "sensitivity", "specificity"),
                              best.method = "youden", transpose = F)

#Selecting the threshold value as probability cut off
```

```
threshold_youden_lr$threshold
```

[1] 0.108166

```
cutoff_lr <- threshold_youden_lr$threshold

#Confusion Matrix using threshold cut off value
pred_logistic <- ifelse(predict(fit_logistic_reg,
                                newdata = test_set,
                                type = "response") > cutoff_lr, 1, 0)

confusionMatrix_lr <- confusionMatrix(data = as.factor(pred_logistic),
                                      reference = as.factor(test_set$DEPRN),
                                      positive = "1")
confusionMatrix_lr
```

Confusion Matrix and Statistics

```
      Reference
```

Prediction 0 1 0 5395 307 1 2810 631

```
           Accuracy : 0.6591
             95% CI : (0.6493, 0.6688)
No Information Rate : 0.8974
P-Value [Acc > NIR] : 1

              Kappa : 0.1514
```

Mcnemar's Test P-Value : <2e-16

```
        Sensitivity : 0.67271
        Specificity : 0.65753
     Pos Pred Value : 0.18338
     Neg Pred Value : 0.94616
         Prevalence : 0.10259
     Detection Rate : 0.06901
```

Detection Prevalence : 0.37635
Balanced Accuracy : 0.66512

```
   'Positive' Class : 1
```

```
compare_results <-data.frame(ML_model = "Logistic Regression",
                      Accuracy = confusionMatrix_lr$overall['Accuracy'],
                      Sensitivity = confusionMatrix_lr$byClass['Sensitivity'],
                      Specificity = confusionMatrix_lr$byClass['Specificity'],
                      Balanced_accuracy = confusionMatrix_lr$byClass['Balanced Accuracy'],
                      AUC = auc_lr)

#Clean up memory
rm(fit_logistic_reg,pred_logistic,threshold_youden_lr,
   roc_lr,confusionMatrix_lr,cutoff_lr, auc_lr)
invisible(gc())
```

3.2 Classification and Regression Trees (CART)

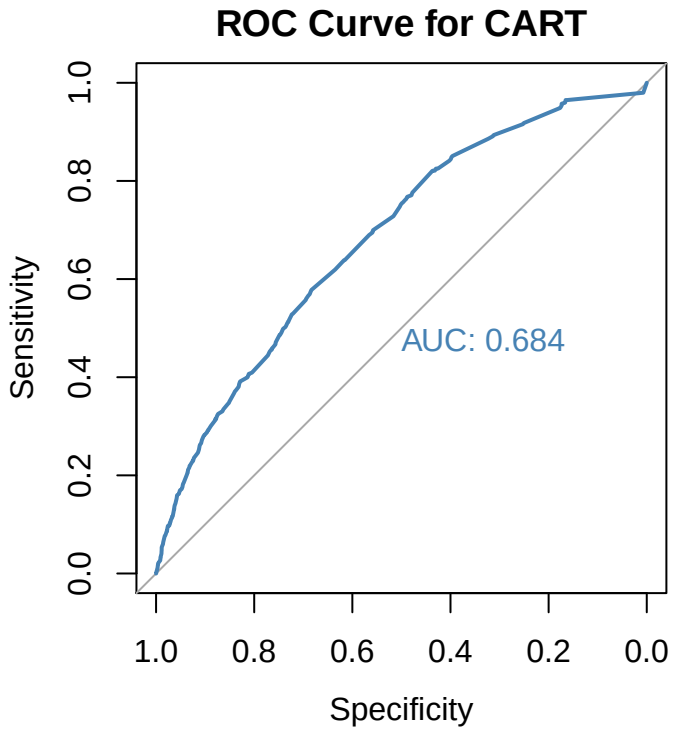
The second machine learning model applied in this project is the **rpart** package in **R** which trains a model using Recursive Partitioning. Initially we tuned the complexity parameter (cp) using different values in the train function of **caret** package.

```
#Classification tree model
set.seed(1000, sample.kind = 'Rounding')
fit_rpart <- rpart(as.factor(DEPRN) ~ .,data = train_set,
                   method = "class",
                   control = rpart.control(minsplit = 20,
                                           cp = 0.00001,
                                          xval = 5,
                                          maxdepth = 30),
                   parms = list(split = "information"))

#ROC graph and area under the curve
y_hat_rpart <- predict(fit_rpart, newdata = test_set,type="prob")[,2]
roc_rpart <- roc(response = as.factor(test_set$DEPRN), predictor = y_hat_rpart)
```

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

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

```
plot(roc_rpart, print.auc=TRUE,
     lwd=2, xlim=c(1,0),
     col="steelblue",
     main="ROC Curve for CART",)
```




```
auc_rpart <- auc(roc_rpart)
auc_rpart
```

Area under the curve: 0.6836

```
# Threshold using youden method
threshold_youden_rpart <- coords(roc_rpart, "best",
                                 ret = c("threshold", "sensitivity", "specificity"),
                                 best.method = "youden", transpose = F)

#Selecting the threshold value as probability cut off
threshold_youden_rpart$threshold
```

[1] 0.09007761

```
cutoff_rpart <- threshold_youden_rpart$threshold

#Confusion Matrix using threshold cut off value
pred_rpart <- ifelse(predict(fit_rpart,
                             newdata = test_set,
                             type = "prob") > cutoff_rpart, 1, 0)[,2]
confusionMatrix_rpart <- confusionMatrix(data = as.factor(pred_rpart),
                                         reference = as.factor(test_set$DEPRN),
                                         positive = "1")
confusionMatrix_rpart
```

Confusion Matrix and Statistics

```
      Reference
```

Prediction 0 1 0 5607 396 1 2598 542

```
           Accuracy : 0.6725
             95% CI : (0.6628, 0.6822)
No Information Rate : 0.8974
P-Value [Acc > NIR] : 1

              Kappa : 0.1281
```

Mcnemar's Test P-Value : <2e-16

```
        Sensitivity : 0.57783
        Specificity : 0.68336
     Pos Pred Value : 0.17261
     Neg Pred Value : 0.93403
         Prevalence : 0.10259
     Detection Rate : 0.05928
```

Detection Prevalence : 0.34343
Balanced Accuracy : 0.63059

```
   'Positive' Class : 1
```

```
compare_results <- compare_results %>%
  add_row(ML_model = "CART",
          Accuracy = confusionMatrix_rpart$overall['Accuracy'],
          Sensitivity = confusionMatrix_rpart$byClass['Sensitivity'],
          Specificity = confusionMatrix_rpart$byClass['Specificity'],
          Balanced_accuracy = confusionMatrix_rpart$byClass['Balanced Accuracy'],
          AUC = auc_rpart)
```

```
#Clean up memory
rm(fit_rpart,y_hat_rpart,pred_rpart,threshold_youden_rpart,
   roc_rpart,confusionMatrix_rpart,cutoff_rpart, auc_rpart)
invisible(gc())
```

3.3 Random Forest

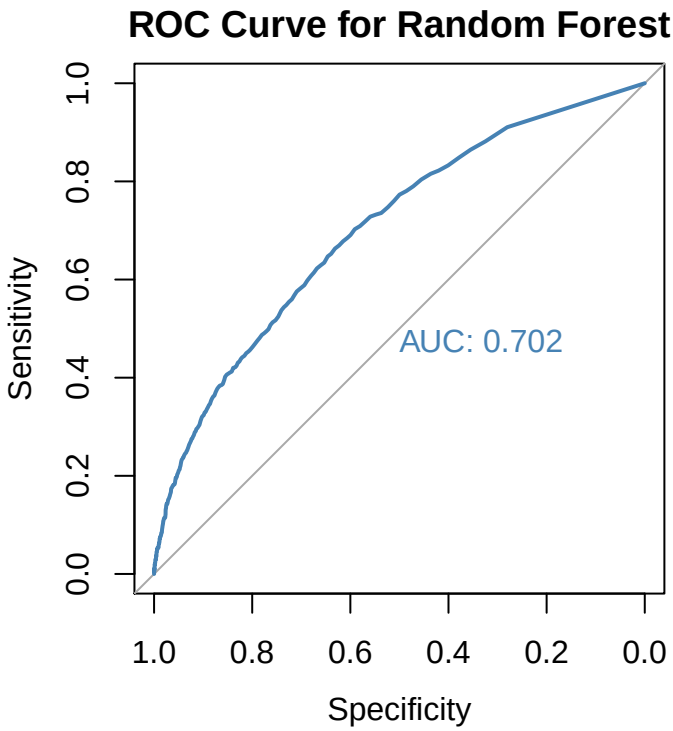
The Random Forest model was implemented using rf method in the caret package of R.

```
#Random Forest model
set.seed(1000, sample.kind = 'Rounding')
fit_rf <- randomForest(as.factor(DEPRN) ~ .,
                       data = train_set,
                       #More trees such as >500 would likely result in
                       #better prediction but would require more powerful
                       #computational resources
                       ntree = 500, mtry = 5, importance = TRUE)


#Make prediciton
y_hat_rf <- predict(fit_rf, test_set, type='prob')[,"1"]

#ROC graph and area under the curve
roc_rf <- roc(response = as.factor(test_set$DEPRN), predictor = y_hat_rf)
```

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

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

```
plot(roc_rf, print.auc=TRUE,
     lwd=2, xlim=c(1,0),
     col="steelblue",
     main="ROC Curve for Random Forest")
```




```
#Extract AUC stats
auc_rf <- auc(roc_rf)
auc_rf
```

Area under the curve: 0.7018

```
# Threshold using youden method
threshold_youden_rf <- coords(roc_rf, "best",
                              ret = c("threshold", "sensitivity", "specificity"),
                              best.method = "youden", transpose = F)

#Selecting the threshold value as probability cut off
threshold_youden_rf$threshold
```

[1] 0.047

```
cutoff_rf <- threshold_youden_rf$threshold

#Confusion Matrix using predictions
pred_rf <- ifelse(predict(fit_rf,
                          newdata = test_set,
                          type = "prob") > cutoff_rf, 1, 0)[,"1"]
confusionMatrix_rf <- confusionMatrix(data = as.factor(pred_rf),
                                      reference = as.factor(test_set$DEPRN),
                                      positive = "1")
confusionMatrix_rf
```

Confusion Matrix and Statistics

Reference

Prediction 0 1 0 5180 316 1 3025 622

```
           Accuracy : 0.6346
             95% CI : (0.6246, 0.6445)
No Information Rate : 0.8974
P-Value [Acc > NIR] : 1

              Kappa : 0.1292
```

Mcnemar's Test P-Value : <2e-16

```
        Sensitivity : 0.66311
        Specificity : 0.63132
     Pos Pred Value : 0.17055
     Neg Pred Value : 0.94250
         Prevalence : 0.10259
     Detection Rate : 0.06803
```

Detection Prevalence : 0.39888
Balanced Accuracy : 0.64722

```
   'Positive' Class : 1
```

```
#Save results in a dataframe
compare_results <- compare_results %>% add_row(ML_model = "Random Forest",
                      Accuracy = confusionMatrix_rf$overall['Accuracy'],
                      Sensitivity = confusionMatrix_rf$byClass['Sensitivity'],
                      Specificity = confusionMatrix_rf$byClass['Specificity'],
                      Balanced_accuracy = confusionMatrix_rf$byClass['Balanced Accuracy'],
                      AUC = auc_rf)

#Clean up memory
rm(fit_rf,y_hat_rf,pred_rf,threshold_youden_rf,
   roc_rf,confusionMatrix_rf,cutoff_rf, auc_rf)
invisible(gc())
```

4 Results and Discussion

A summarized description of the performance of four Machine Learning Models used in this project is listed in the following comparison table.

```
#Print out summary of results
kable(compare_results) %>%
  kable_styling(full_width = F)%>%
  row_spec(0, bold = T, color = "black", background = "#EBF2F6")
```

ML_model	Accuracy	Sensitivity	Specificity	Balanced_accuracy	AUC
Logistic Regression	0.6590835	0.6727079	0.6575259	0.6651169	0.7258161
CART	0.6725364	0.5778252	0.6833638	0.6305945	0.6835707
Random Forest	0.6345838	0.6631130	0.6313224	0.6472177	0.7017800

The result shows that AUC statistics were similar in all the models with Logistic Regression having the highest AUC. There were slight differences in other performance measures. For example, CART had the highest accuracy and specificity but lowest sensitivity. The findings indicate that there is a trade-off of precision in correctly detecting positive cases of depression and correctly ruling out depression. Therefore,

care should be taken in selecting the optimally tuned model depending on the context in which the findings may be applied.

5 Conclusions

In this project, several predictive models for depression were built based on LR, CART, and RF using routinely available data from the electronic health records in a large representative data from the US National Ambulatory Medical Care Survey. No single model outperformed the other in term of all the performance measures. Therefore, optimum model will depend on the health care context. Fure research with more features may be able to improve the predictive ability of current models.

The findings further support the power of large electronic health data and promising applications of machine learning algorithms to offer new insights to inform clinical and public health policy.

One of the major criticisms of machine learning and artificial intelligence in medicine lies in the fact that the predictors are selected within a 'black box' without proper clinical relevance. To address this, extensive literature was consulted to select potentially relevant predictors of depression in this project. Therefore, the clinicians and public health policy makers will hopefully find the results from the machine learning algorithms used in this project relevant to clinical decision making.

6 References

[1] Patel V, Chisholm D, Parikh R, et al. Addressing the burden of mental, neurological, and substance use disorders: key messages from Disease Control Priorities. The Lancet. 2016;387(10028):1672-85.

[2] Herrman H, Kieling C, McGorry P, et al. Reducing the global burden of depression: a Lancet–World Psychiatric Association Commission. The Lancet. 2019;393(10189):e42-3.

[3] Brody DJ, Pratt LA, Hughes JP. Prevalence of Depression Among Adults Aged 20 and Over: United States, 2013-2016. NCHS Data Brief. 2018;(303):1-8.

[4] National Institute of Mental Health. Major depression. February 2019. URL: https://www.nimh.nih.gov /health/statistics/major-depression.shtml

[5] Freeman A, Tyrovolas S, Koyanagi A, et al. The role of socio-economic status in depression: results from the COURAGE (aging survey in Europe). BMC Public Health. 2016;16(1):1098.

[6] World Health Organization. Depression: Key Facts. December 4, 2019. URL: https://www.who.int/news-room/fact-sheets/detail/depression

[7] Kohn R, Saxena S, Levav I, Saraceno B. The treatment gap in mental health care. Bulletin of the World Health Organization. 2004;82(11):858–66.

[8] Halfin A. Depression: the benefits of early and appropriate treatment. American Journal of Managed Care. 2007;13(4 Suppl):S92-7.

[9] Chen JH, Asch SM. Machine learning and prediction in medicine—beyond the peak of inflated expectations. New England Journal of Medicine. 2017;376(26):2507.

[10] Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. New England Journal of Medicine. 2016;375(13):1216.

[11] McKinney, S.M., Sieniek, M., Godbole, V. et al. International evaluation of an AI system for breast cancer screening. Nature. 2020;577:89–94.

[12] National Health Care Surveys Registry. 2014 NAMCS MICRO-DATA FILE DOCUMENTATION. URL: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NAMCS/doc2014.pdf.

[13] Ladapo JA, Larochelle MR, Chen A, et al. Physician prescribing of opioids to patients at increased risk of overdose from benzodiazepine use in the United States. JAMA psychiatry. 2018;75(6):623-30.

[14] Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biometrical Journal: Journal of Mathematical Methods in Biosciences. 2005;47(4):458-72.