Data Exploration for Alzheimer Patients Dataset

Phase -2

Data Set Information:

The Excel spreadsheet Alzheimer.csv contains one sheet named Alzheimer, which is data attempting to explain whether a patient has Alzheimer's Disease. These are data from a sample of 336 employees and consists of 9 variables for each patient. These are:

1) Dementia-Outcome variable-patient diagnosis
2) Gender-Female=0 and Male=1
3) Age-Age of patient (in years)
4) Education-Years of Education
5) SES-Socioeconomic Status 1=Low and 5=High
6) MMSE-Mini mental state examination score
7) CDR-Clinical Dementia Rating
8) eTIV-estimated total intracranial volume
9) nWBV-Normalize whole brain volume
10) ASF-Atlas Scaling Factor

Developed a Logistic Regression model to classify the Dementia event from the other variables.

```
library(MASS)
library(plyr)
library(readr)
library(tidyr)
library(broom)
library(gtsummary)
library(dplyr)     # data wrangling
library(ggplot2)   # plotting
library(rsample)   # training and testing splitting
library(caret)     # for logistic regression modeling and prediction outputs
library(vip)       # variable importance

setwd("C:/Users/DELL/Desktop/MS Assignments/Sem1/Data_Stats/Assignmnet8")
#Read in Datasets
Alz <- read.csv(file="alzheimer-1.csv")

#Check data is read in correctly
View(Alz)

head(Alz)

#Check Variable Names
names(Alz)


#Check Data structure
str(Alz)


#Check Missing Data
sum(is.na(Alz))

#Treat Missing Values

#Listwise Deletion
Alz_new <- na.omit(Alz)

#Check new data has no missing data
sum(is.na(Alz_new))
View(Alz_new)
head(Alz_new)

########################################
Alz_new$Dementia<- revalue(Alz_new$Dementia,c("Alzheimer"=0, "No Alzheimer"=1))
Alz_new$Dementia<- as.factor(Alz_new$Dementia)
View(Alz_new)
```

```
> names(Alz)
 [1] "Dementia" "Gender"   "Age"      "EDUC"     "SES"      "MMSE"     "CDR"      "eTIV"
 [9] "nWBV"     "ASF"
> #Check Data structure
> str(Alz)
'data.frame':   1008 obs. of  10 variables:
 $ Dementia: chr  "No Alzheimer" "No Alzheimer" "Alzheimer" "Alzheimer" ...
 $ Gender  : int  1 1 1 1 1 0 0 1 1 1 ...
 $ Age     : int  87 88 75 76 80 88 90 80 83 85 ...
 $ EDUC    : int  14 14 12 12 12 18 18 12 12 12 ...
 $ SES     : int  2 2 NA NA NA 3 3 4 4 4 ...
 $ MMSE    : int  27 30 23 28 22 28 27 28 29 30 ...
 $ CDR     : num  0 0 0.5 0.5 0.5 0 0 0 0.5 0 ...
 $ eTIV    : int  1987 2004 1678 1738 1698 1215 1200 1689 1701 1699 ...
 $ nWBV    : num  0.696 0.681 0.736 0.713 0.701 0.71 0.718 0.712 0.711 0.705 ...
 $ ASF     : num  0.883 0.876 1.046 1.01 1.034 ...
> #Check Missing Data
> sum(is.na(Alz))
[1] 63
> #Listwise Deletion
> Alz_new <- na.omit(Alz)
> #Check new data has no missing data
> sum(is.na(Alz_new))
[1] 0
> View(Alz_new)
> head(Alz_new)
      Dementia Gender Age EDUC SES MMSE CDR eTIV  nWBV   ASF
1 No Alzheimer      1  87   14   2   27 0.0 1987 0.696 0.883
2 No Alzheimer      1  88   14   2   30 0.0 2004 0.681 0.876
6 No Alzheimer      0  88   18   3   28 0.0 1215 0.710 1.444
7 No Alzheimer      0  90   18   3   27 0.0 1200 0.718 1.462
8 No Alzheimer      1  80   12   4   28 0.0 1689 0.712 1.039
9 No Alzheimer      1  83   12   4   29 0.5 1701 0.711 1.032
> ##########################################
> Alz_new$Dementia<- revalue(Alz_new$Dementia,c("Alzheimer"=0, "No Alzheimer"=1))
```

a) Created a logistic regression model and explain the significant odds ratios in terms of Dementia.

```
48  #a)
49  set.seed(123)  # use a set seed point for reproducibility
50  split <- initial_split(Alz_new, prop = .7, strata = "Dementia")
51  train <- training(split)
52  test  <- testing(split)
53
54  #Logistic Regression
55
56  #For explaining dependent variable
57
58  Alz_new$Dementia <- as.factor(Alz_new$Dementia)
59
60  log_reg <- glm(
61    Dementia ~ Gender + Age + EDUC + SES + MMSE + eTIV +CDR+ nWBV + ASF   ,
62    family = "binomial",
63    data = Alz_new
64  )
65
66  summary(log_reg)
67
68  tidy(log_reg)
69
70  #Coefficients in exponential form
71  log_reg %>%
72    gtsummary::tbl_regression(exp = TRUE)
```

From the Odds Ratio we can know that if OR is greater than 1, than the effect of the parameter will be strong and will have positive effect. Whereas if the OR is less than 1, that means the event, or the parameter does not have much effect.

So, from the below output we get the following results:

1. Age, SES, MMSE have a positive effect on the Dementia parameter.

2. ASF rating has an extremely high value of odds ratio and that shows and extremely strong positive impact on dementia.
3. Gender, CDR, EDUC have values less than 1, hence do not have a positive impact on the Dementia parameter.

Characteristic	OR[1]
Gender	0.00
Age	647
EDUC	0.02
SES	3,424,606
MMSE	600
eTIV	1.53
CDR	0.00
nWBV	Inf
ASF	222,226,539,347,334,259,248,466,422,288,840,462,080,008,682,868,004,026,246,622,824,426,226,840,242,666,442,240,026,848,882,280,082,028,284

[1] OR = Odds Ratio, CI = Confidence Interval

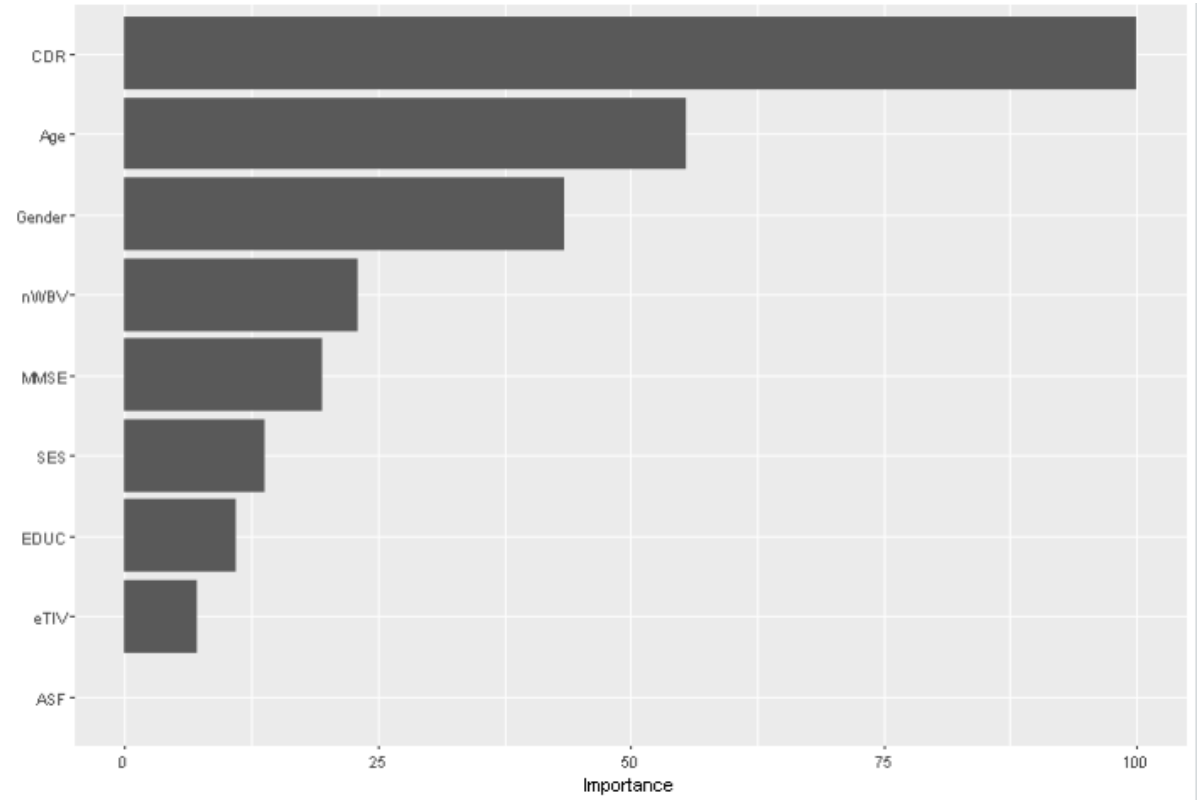
b) Created a confusion matrix and explained how the model is classifying who has Dementia.

```
#b)

train$Dementia <- as.factor(train$Dementia)

#For Predicting dependent variable
log_reg = train(
  form = Dementia ~ Gender + Age + EDUC + SES + MMSE + CDR + eTIV + nWBV + ASF,
  data = train,
  method = "glm",
  family = "binomial"
)

pred <- predict(log_reg, test)
pred


#Confusion Matrix
confusionMatrix(pred, as.factor(test$Dementia))
vip(log_reg, num_features = 10)
```

```
> pred
  [1] 1 1 1 1 0 0 1 1 1 0 0 0 1 1 0 1 1 1 0 0 1 1 0 0 1 0 0 0 1 1 0 0 1 1 1 1 1 0 1 1 0 0 0 1 1 1
 [47] 1 1 0 0 0 0 1 0 0 1 1 1 1 1 0 0 1 1 0 0 0 0 1 0 1 1 1 0 1 1 1 0 1 1 0 0 1 1 1 1 1 0 1 1 1 0 1
 [93] 0 0 1 1 1 0 0 1 1 1 1 0 0 0 0 1 1 0 0 1 1 1 1 1 1 0 0 0 0 1 0 1 0 1 0 1 1 1 1 0 0 0 0 1 1 1 1 1 1 1 0 1
[139] 1 0 0 1 1 0 0 1 1 1 1 1 1 1 1 1 0 0 1 1 0 0 1 1 1 1 1 1 0 0 1 0 1 1 1 0 1 1 0 1 1 1 0 1 0 0 0 0 1
[185] 1 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 1 0 1 1 1 1 1 0 0 0 0 0 0 1 0 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1
[231] 0 1 1 1 1 1 0 0 0 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 1 0 0 1 1 1 1 0 1 1 0 0 0 0 1 1 0 0 1 1 1 1 1
[277] 0 0 1 1 1 1 0 0 1 1
Levels: 0 1
> #Confusion Matrix
> confusionMatrix(pred, as.factor(test$Dementia))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 115   0
         1   0 171

               Accuracy : 1
                 95% CI : (0.9872, 1)
    No Information Rate : 0.5979
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.4021
         Detection Rate : 0.4021
   Detection Prevalence : 0.4021
      Balanced Accuracy : 1.0000

       'Positive' Class : 0

> vip(log_reg, num_features = 10)
```

From the above image of confusion matrix, we can find that we have 115 true positives and 171 true negatives. The accuracy is 1 which depicts 100% for Dementia prediction, showing that the model

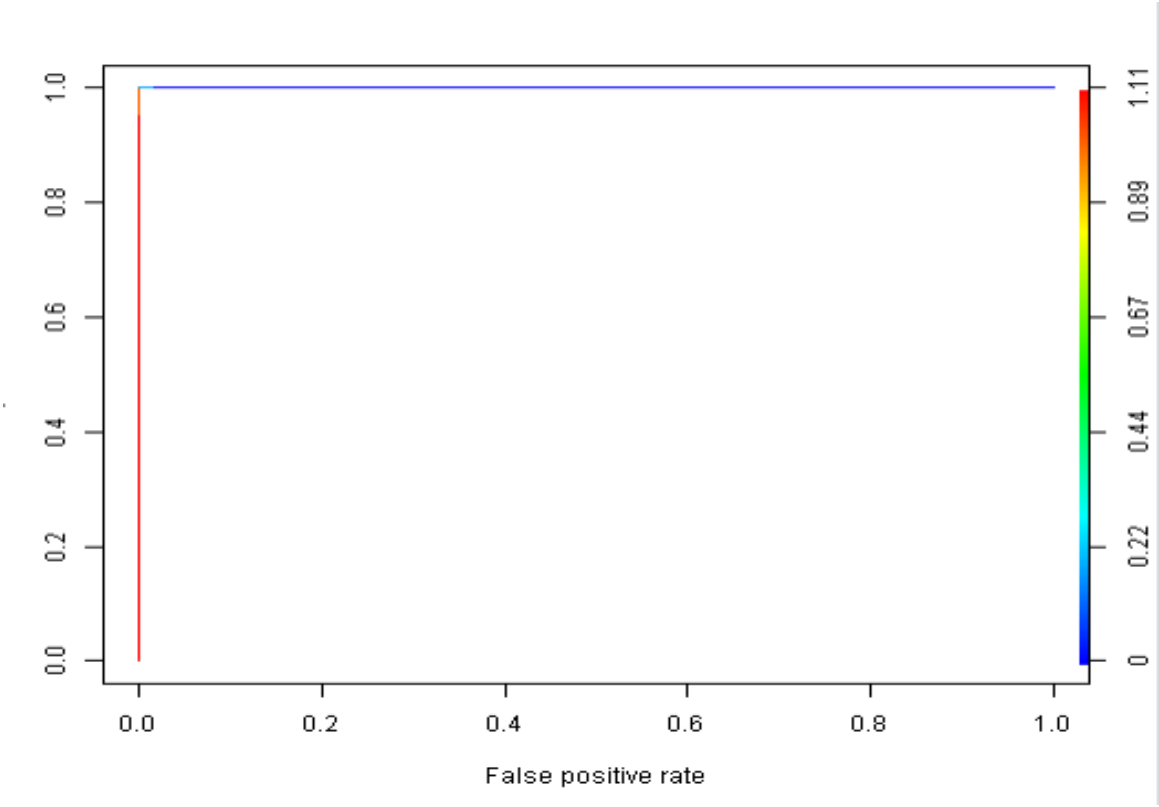
correctly predicts if a patient has Dementia or not. The model's sensitivity is also 100% along with the Specificity that indicates that there are no false negatives.

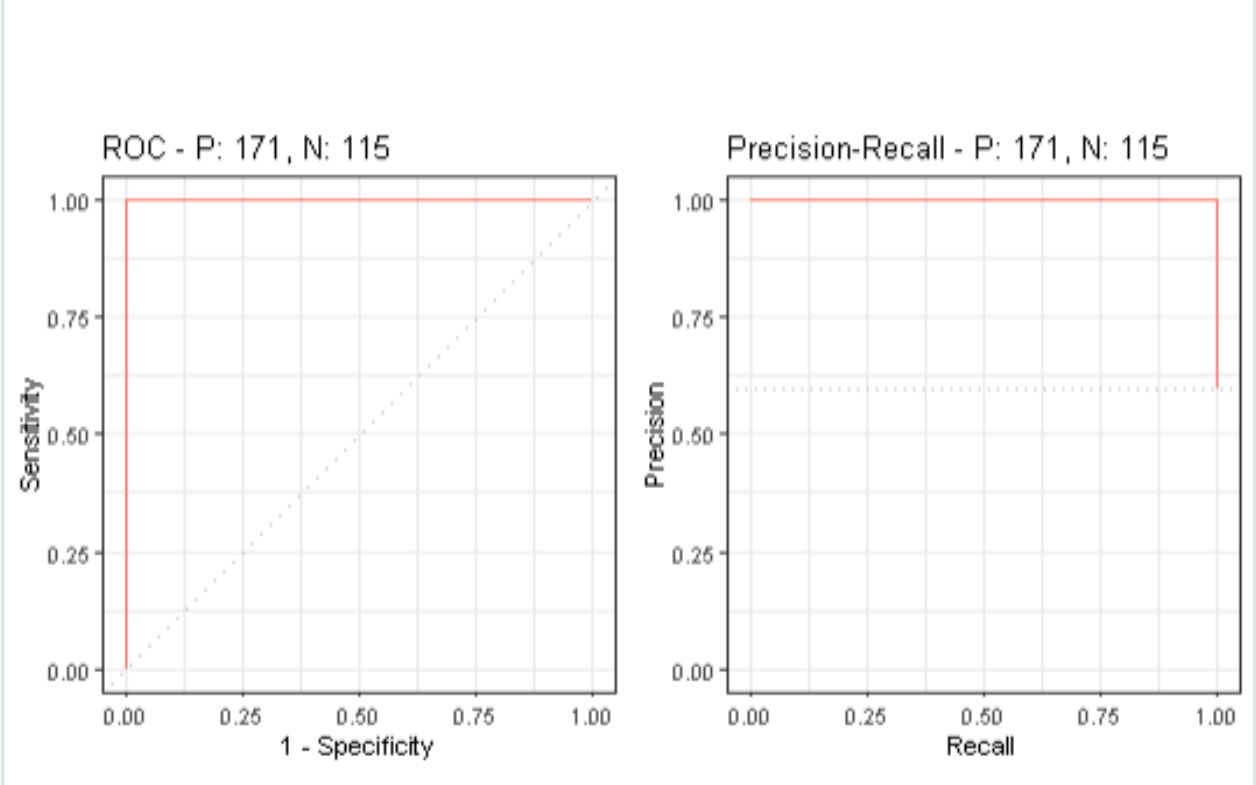


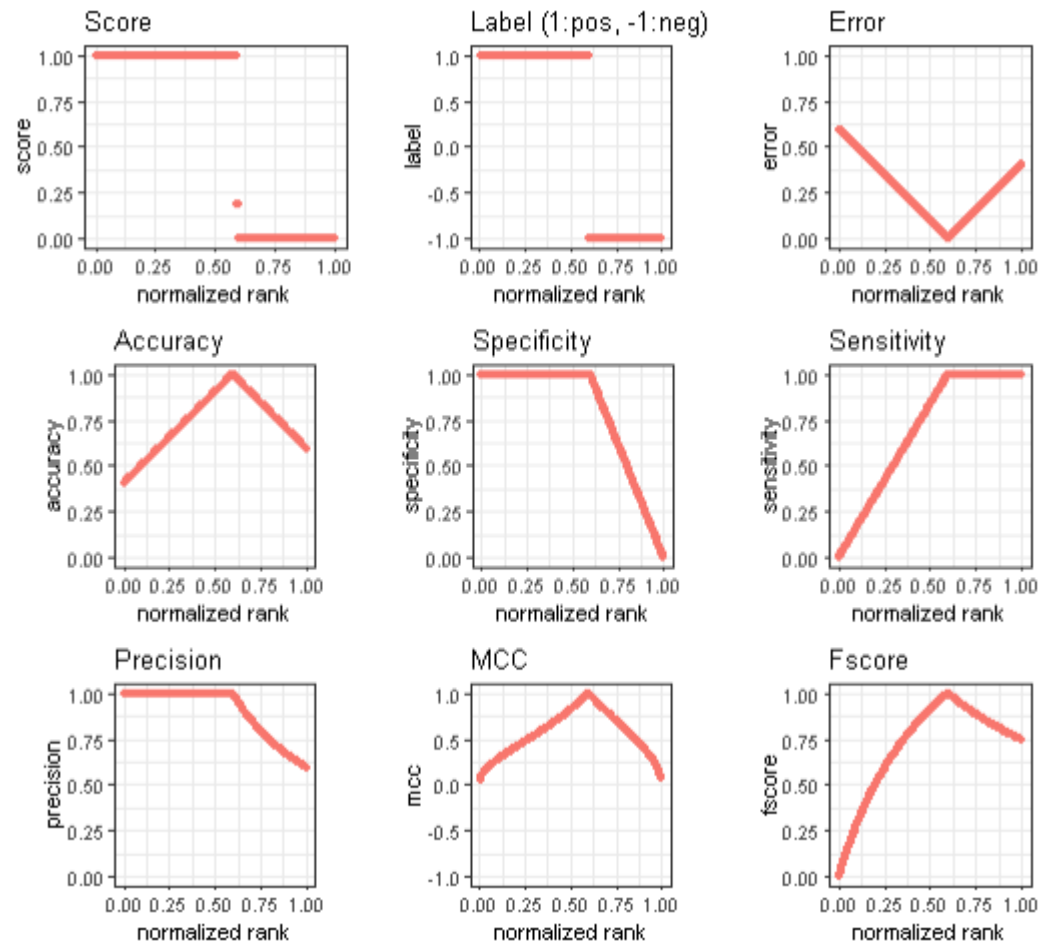
We can also see from the above graph that the importance of CDR is 100%.

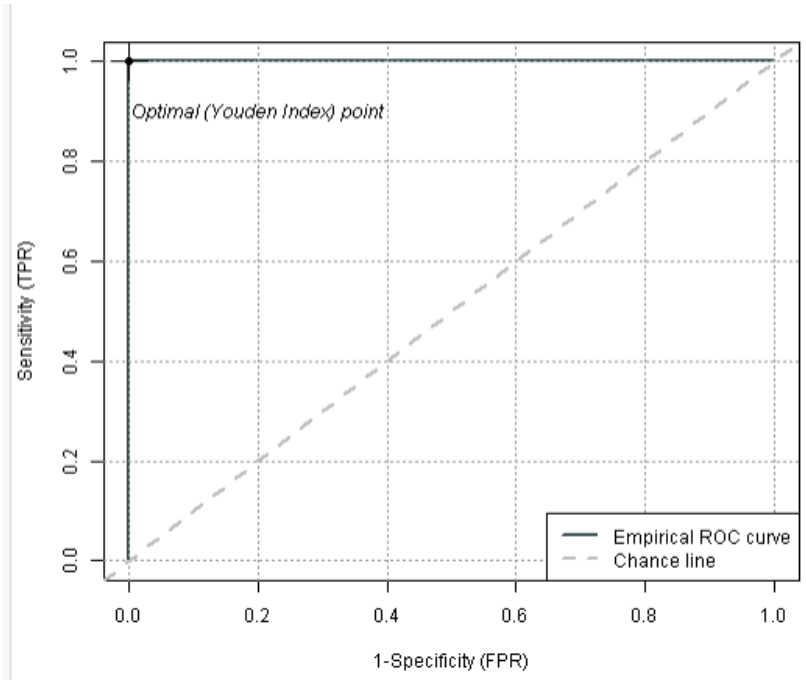
c) Created an ROC curve and calculate the c-statistic (auc). Information about the model.

```
#c)

#ROC Curves

log_reg_train <- glm(Dementia ~ Gender + EDUC + Age + SES + MMSE + CDR+eTIV,
                     data = train,  family = "binomial")
library(ROCR)

log_reg_test_prob <- log_reg_train %>% predict(test, type = "response")
log_reg_test_prob

preds <- prediction(as.numeric(log_reg_test_prob), test$Dementia)

perf <- performance(preds,"tpr","fpr")
plot(perf,colorize=TRUE)

library(precrec)
precrec_obj <- evalmod(scores = log_reg_test_prob, labels = test$Dementia)
autoplot(precrec_obj)

## Get AUCs
sm_aucs <- auc(precrec_obj)
## Shows AUCs
sm_aucs

precrec_obj2 <- evalmod(scores = log_reg_test_prob, labels = test$Dementia, mode="basic")
autoplot(precrec_obj2)

library(ROCit)

ROCit_obj <- rocit(score=log_reg_test_prob,class=test$Dementia)
plot(ROCit_obj)

summary(ROCit_obj)

measure <- measureit(score = log_reg_test_prob, class = test$Dementia,
                     measure = c("ACC", "MIS", "SENS", "SPEC", "PREC", "REC","PPV","NPV", "FSCR"))
measure
```


0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
1.11
0.89
0.67
0.44
0.22
0


ROC - P: 171, N: 115
Sensitivity
1 - Specificity
0.00
0.25
0.50
0.75
1.00
Precision-Recall - P: 171, N: 115
Precision
Recall





From the above results we find following observations:

1. From the first graph we find that the true positive rate is 1 whereas false positive and negative are 0.

2. The sensitivity and specificity shown in the graphs are optimal i.e 100% for both ROC curve and precision.
3. Thus, we can accurately predict the dementia variable from the parameters.
4. Multiple ROC's are used to compare the AUC. Models having higher AUC are recognized to be more better models than those of lower. So comparing the ROC and using AUC can help us determine the methods better.
5. Thus our ROC is 100% from the above graph which we plotted between TPR vs FPR.

d) Differences between the information in part a and part b?

With Part A, we were able to determine the which factors affect the most for the dementia parameter and can help us predict the patients with Dementia correctly. We could see with the help of OR tat how much impact does each variable have on dementia. Hence, we could find the positive and negative impact of variable on dementia.

Whereas is Part B, we find the accuracy of the model to predict the dementia of the patient. The confusion matrix helps us in showing the model's performance. We can also find the true positive and negatives and false positives and negatives with the help of part B.

e) Information on Model difference from the linear discriminate analysis in Phase 1?

The Logistic regression determines the accuracy of an event to happen whereas the Linear Discriminant Analysis determines where there is correctness in the event occurring. The LDA can be used to analyse large amount of data and then be plotted in graphs whereas Logistic Regression is used for categorical values which can be divided into groups. We can find sensitivity in the Logistic regression, and which tends to be not much sensitive to large amount of scattered data unlike LDA. Thus, logistic regression is based on maximizing the likelihood unlike the LDA. Also, in this case logistic regression performed with better accuracy than LDA.