Stats141XP - HW 2 (Charles' Part)

Charles Liu (304804942)

4/14/2021

Part Two

1. Using the diabeticsub data posted in the homework folder on week two, predict the odds of diabetic type II (Diabetes), as a function of hypertension, age, family history of diabetic, and the interaction effect between family history of diabetic and age.
2. Using the library caTools , 65% of the data as testing, and median of predicted scores as threshold, estimate the accuracy of the model.
3. Using mean of predicted scores as threshold, re-calculate the accuracy of the model. Compare the accuracy based on using mean and median as the threshold.
4. Create a ROC curve for estimating the accuracy of the model created for the prediction diabetes. Use 65% of the data as testing.
5. Create a contingency table and estimate the accuracy based on the best cut-off estimated based on the ROC curve.
6. Draw a Roc curve showing true positive and false positive rate and explain what it shows.
7. Perform a five-fold cross validation and explain the findings. Use 70% of the data as testing. Is the accuracy resulting from five-fold cross validation better than the one resulting from the confusion matrix created based on the median and mean cutoff from the ROC Curve?
8. Define sensitivity and specificity and show how you can calculate them for the confusion matrix created in part seven. Discuss whether the model does better with respect to sensitivity or specificity?

Loading Necessary Packages

```r
library(readr)
library(caTools)
library(MASS)
library(pROC)
library(ROCR)
library(car)
library(effects)
```

Loading Necessary Data

```r
setwd(getwd())
d <- read_csv("diabeticsub.csv")
```

Data Set-up

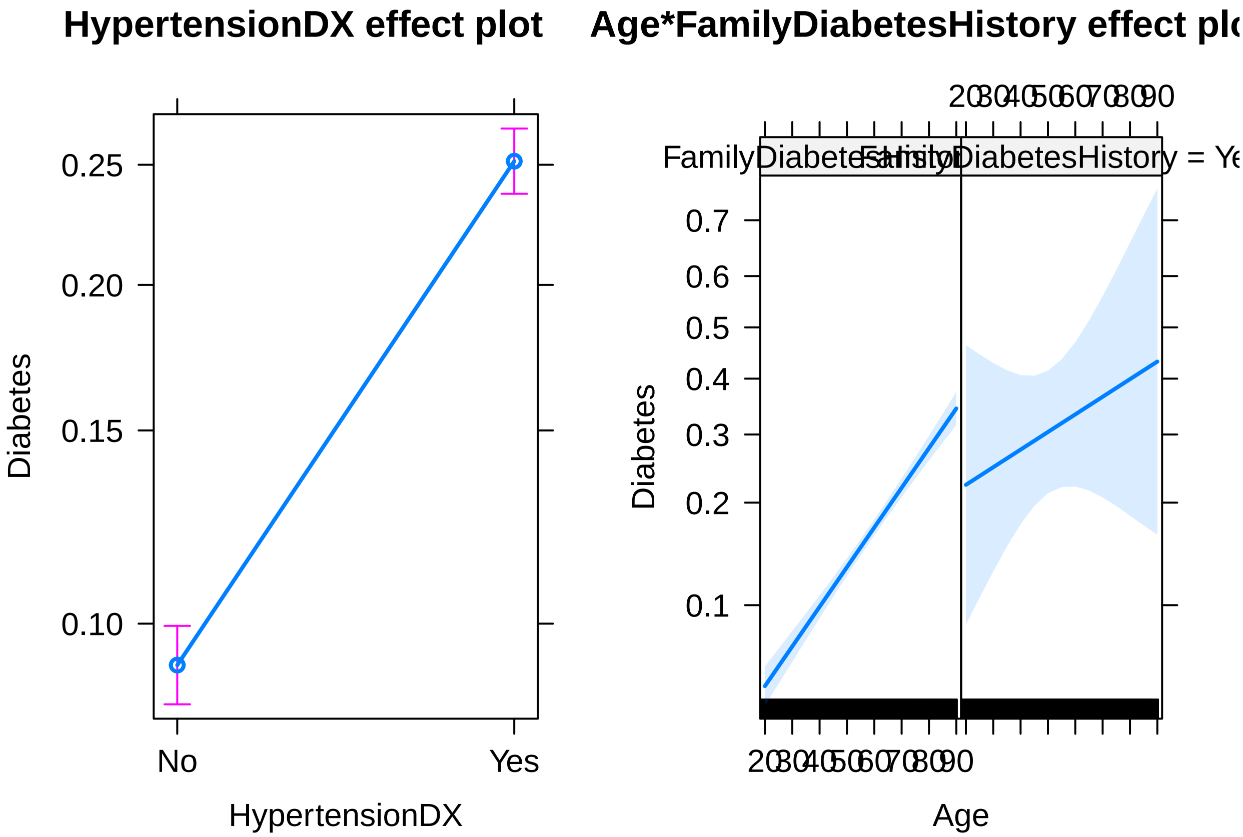
```
# setting up as.factor(...)
d$Diabetes <- as.factor(d$Diabetes)
d$FamilyDiabetesHistory <- as.factor(d$FamilyDiabetesHistory)
```

Q1

```
m1 <- glm(Diabetes ~
               HypertensionDX +
               Age * FamilyDiabetesHistory,
             data = d, family = "binomial")
summary(m1)
```

```
##
## Call:
## glm(formula = Diabetes ~ HypertensionDX + Age * FamilyDiabetesHistory,
##     family = "binomial", data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2909  -0.6877  -0.4101  -0.2889   2.5877
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -3.97103    0.11354 -34.976   <2e-16 ***
## HypertensionDXYes             1.20645    0.06152  19.611   <2e-16 ***
## Age                           0.03137    0.00189  16.602   <2e-16 ***
## FamilyDiabetesHistoryYes      1.94088    0.87990   2.206   0.0274 *
## Age:FamilyDiabetesHistoryYes -0.01744    0.01665  -1.048   0.2947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9422.3  on 9947  degrees of freedom
## Residual deviance: 8270.9  on 9943  degrees of freedom
## AIC: 8280.9
##
## Number of Fisher Scoring iterations: 5
```

```
# Check for Interaction Effect
# For every increase in Age, we can see there is an increase in likely of having Diabetes. It is simila
# HOWEVER, this is NOT statistically significant!
plot(allEffects(m1), ask = FALSE)
```



```r
# Interpretation of Age
exp(m1$coefficients[3] * 10) # Age every 10 years
```

```
##      Age
## 1.368518
```

ANSWER: We can see here that the odds of people with Hypertension (*HypertensionDXYes*) are approximately 20.65% more likely to experience Diabetes. As for Age, we needed to adjust the variable by finding the exponential and multiplying it by 10 (for every 10 years of Age). After the adjustments, we see that the odds of people every 10 years of Age increase will be 36.85% more likely to have Diabetes. The odds of people with a Family History of having Diabetes approximately (*FamilyDiabetesHistoryYes*) is approximately 94.09% more likely to experience Diabetes in their life. Finally, we see that Age and people with a Family History of Diabetes interaction effect is not statistically significant.

Q2

```r
split <- sample.split(d$Diabetes, SplitRatio = 0.65)
train <- subset(d, split = TRUE)
test <- subset(d, split = FALSE)
p <- predict(m1, newdata = test, type = "response")
summary(p)[3] # using the Median as the threshold of 0.1336525
```

```
##    Median
## 0.1336525
```

```
t1 <- table(d$Diabetes, p > summary(p)[3])
sum(diag(t1))/sum(t1) # accuracy
```

```
## [1] 0.6142943
```

ANSWER: We can see that using the median as threshold, we have an accuracy of approximately 61.43% for our model.

Q3

```
summary(p)[4] # using the Mean as the threshold of 0.1814435
```

```
##      Mean
## 0.1814435
```

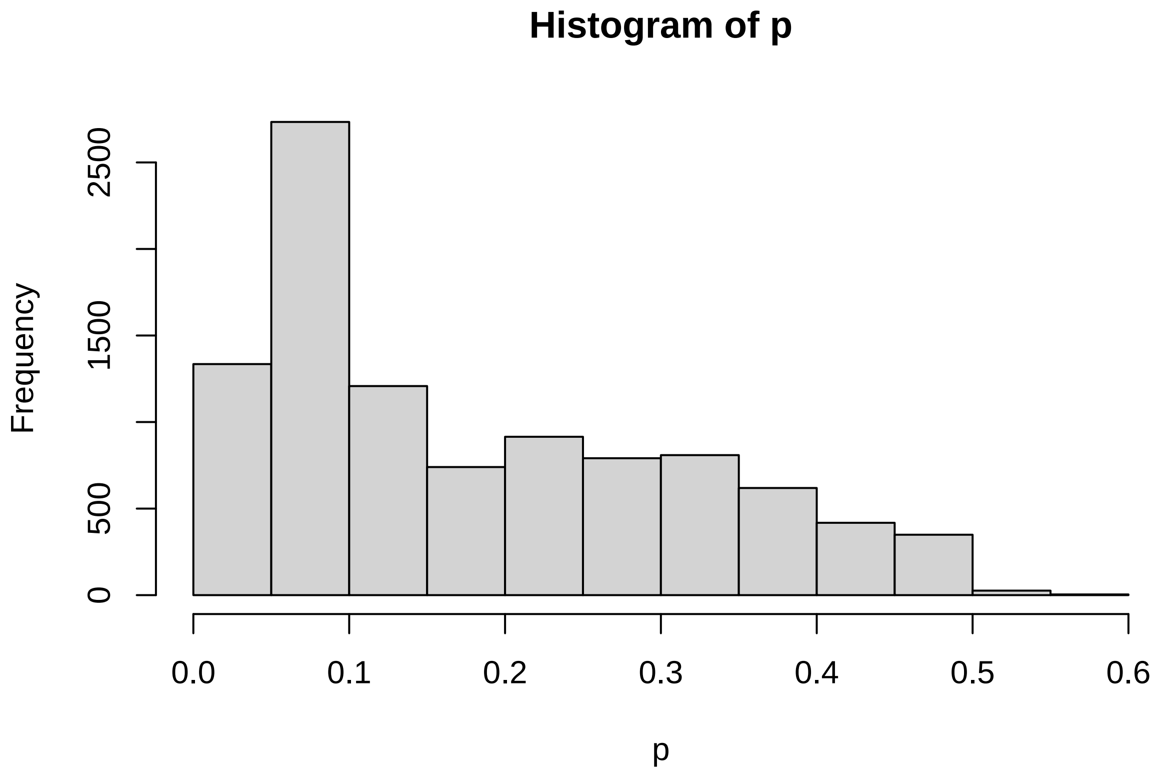
```
t2 <- table(d$Diabetes, p > summary(p)[4])
sum(diag(t2))/sum(t2) # accuracy
```

```
## [1] 0.6640531
```

ANSWER: We can see that using the mean as threshold, we have an accuracy of approximately 66.41% for our model. By comparison, we can see that using the mean as the threshold will give us about a 5% higher accuracy than using the median as a threshold.

Q4

```
pred <- prediction(p, d$Diabetes) # still using 65% as testing
hist(p) # most fall behind 0.5 on the histogram (potential cut-off)
```

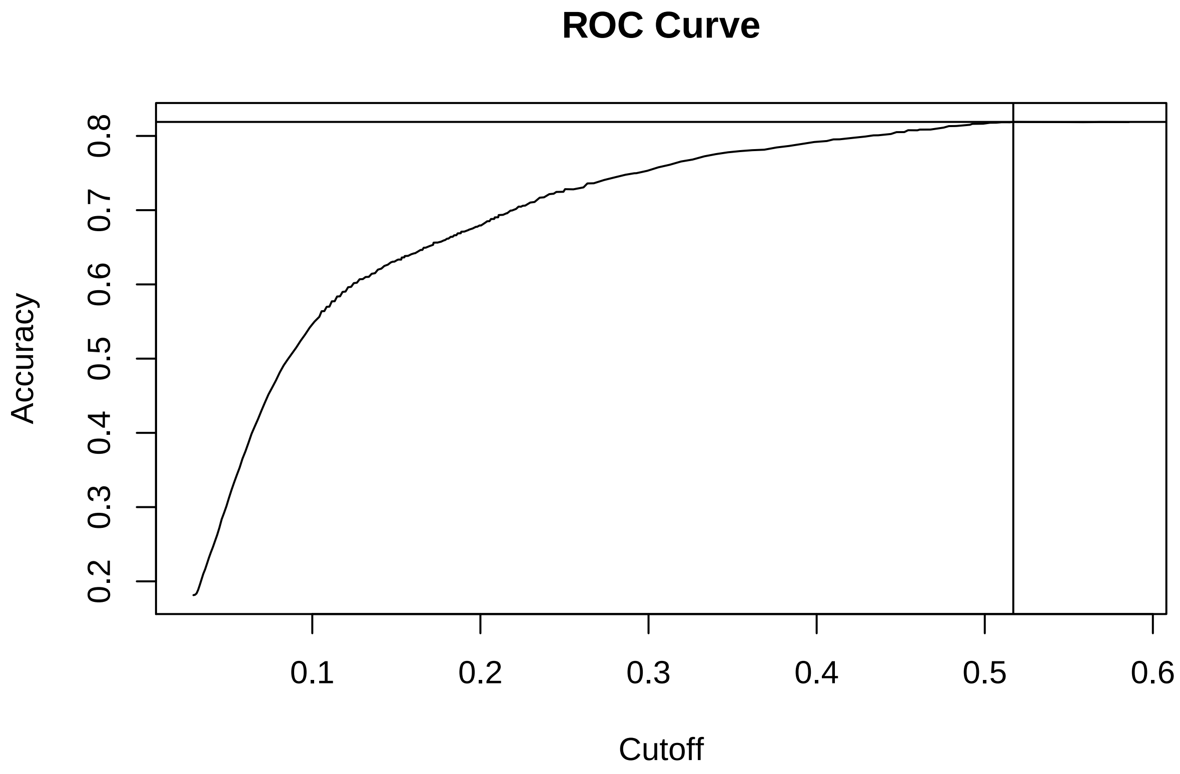


```r
eval <- performance(pred, "acc")

# Estimation of the cutoff that creates the maximum accuracy
max <- which.max(slot(eval, "y.values")[[1]])
acc <- slot(eval, "y.values")[[1]][max]
cut <- slot(eval, "x.values")[[1]][max]
print(c(Accuracy=acc, Cutoff = cut))
```

```
##     Accuracy Cutoff.2849
##    0.8188581   0.5169476
```

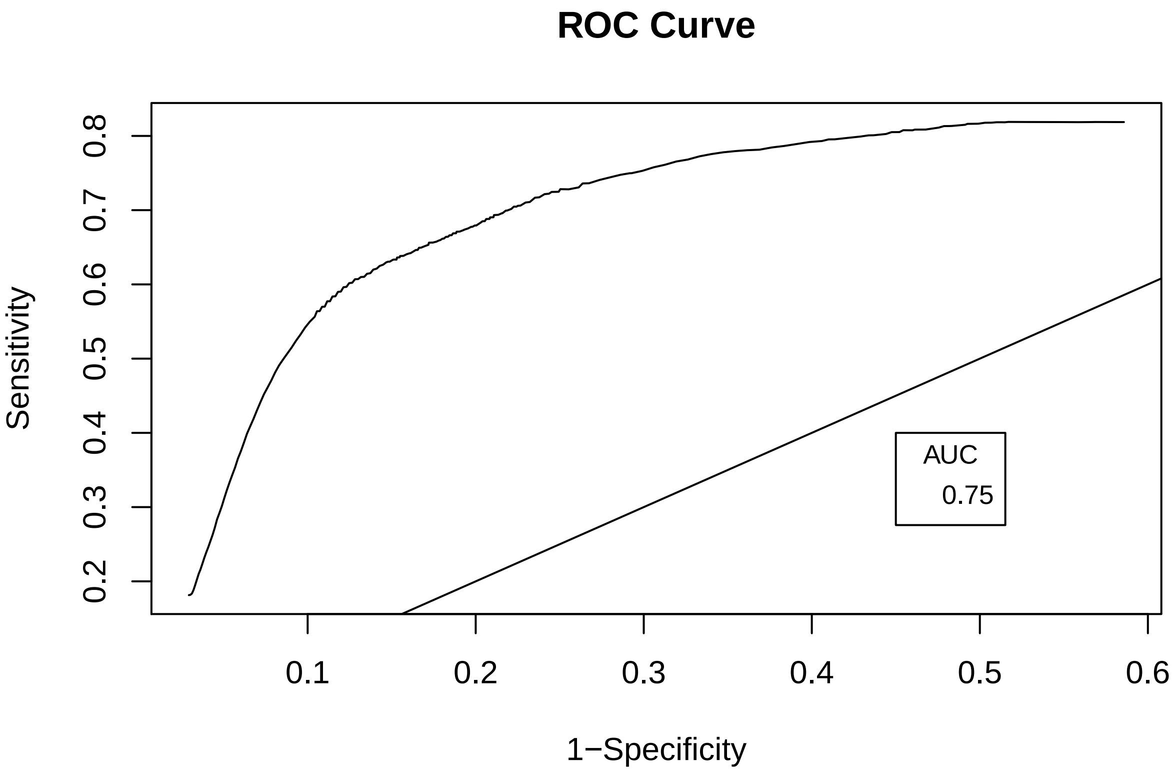
```r
# Plot ROC Curve w/ accuracy & cut-off
plot(eval, main = "ROC Curve")
abline(h=0.8188581,v=0.5169476)
```



```r
# Find AUC
auc <- performance(pred, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,2)
auc
```

```
## [1] 0.75
```

```r
# Plot AUC and ROC Curve
plot(eval, main="ROC Curve", ylab="Sensitivity", xlab="1-Specificity")
abline(0,1)
legend(0.45,0.4,auc,title="AUC",cex=0.8)
```



ANSWER: Using the ROC Curve and plotting it, we can see that the AUC is 75% to our model. We have an accuracy of approximately 81.89% to our model with a cut-off of 51.68%.