Logistic Regression and Correlation Matrix

Ridhwan Firdaus bin Mohd Jata

3/11/2021

library(caTools)

## Warning: package 'caTools' was built under R version 3.6.3

##Dataset In this investigation, we will be using the diabetes dataset. Below is the structure of the dataset

df = read.csv('diabetes.csv', header = TRUE)  
str(df)

## 'data.frame': 768 obs. of 9 variables:  
## $ Pregnancies : int 6 1 8 1 0 5 3 10 2 8 ...  
## $ Glucose : int 148 85 183 89 137 116 78 115 197 125 ...  
## $ BloodPressure : int 72 66 64 66 40 74 50 0 70 96 ...  
## $ SkinThickness : int 35 29 0 23 35 0 32 0 45 0 ...  
## $ Insulin : int 0 0 0 94 168 0 88 0 543 0 ...  
## $ BMI : num 33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...  
## $ DiabetesPedigreeFunction: num 0.627 0.351 0.672 0.167 2.288 ...  
## $ Age : int 50 31 32 21 33 30 26 29 53 54 ...  
## $ Outcome : int 1 0 1 0 1 0 1 0 1 1 ...

head(df)

## Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## 1 6 148 72 35 0 33.6  
## 2 1 85 66 29 0 26.6  
## 3 8 183 64 0 0 23.3  
## 4 1 89 66 23 94 28.1  
## 5 0 137 40 35 168 43.1  
## 6 5 116 74 0 0 25.6  
## DiabetesPedigreeFunction Age Outcome  
## 1 0.627 50 1  
## 2 0.351 31 0  
## 3 0.672 32 1  
## 4 0.167 21 0  
## 5 2.288 33 1  
## 6 0.201 30 0

## Creating Age category column

We will classify all ages into certain category column

df$AgeCat <- ifelse(df$Age < 21, "<21",   
 ifelse((df$Age>=21) & (df$Age<=25), "21-25",   
 ifelse((df$Age>25) & (df$Age<=30), "25-30",  
 ifelse((df$Age>30) & (df$Age<=35), "30-35",  
 ifelse((df$Age>35) & (df$Age<=40), "35-40",  
 ifelse((df$Age>40) & (df$Age<=50), "40-50",  
 ifelse((df$Age>50) & (df$Age<=60), "50-60",">60")))))))  
df$AgeCat <- factor(df$AgeCat, levels = c('<21','21-25','25-30','30-35','35-40','40-50','50-60','>60'))  
table(df$AgeCat)

##   
## <21 21-25 25-30 30-35 35-40 40-50 50-60 >60   
## 0 267 150 81 76 113 54 27

head(df)

## Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## 1 6 148 72 35 0 33.6  
## 2 1 85 66 29 0 26.6  
## 3 8 183 64 0 0 23.3  
## 4 1 89 66 23 94 28.1  
## 5 0 137 40 35 168 43.1  
## 6 5 116 74 0 0 25.6  
## DiabetesPedigreeFunction Age Outcome AgeCat  
## 1 0.627 50 1 40-50  
## 2 0.351 31 0 30-35  
## 3 0.672 32 1 30-35  
## 4 0.167 21 0 21-25  
## 5 2.288 33 1 30-35  
## 6 0.201 30 0 25-30

## Spliting dataset into train and test dataset

require(caTools)  
set.seed(3)  
split = sample.split(df$Outcome,SplitRatio = 0.75)  
train = subset(df, split== TRUE)  
test = subset(df, split== FALSE)  
head(train)

## Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## 1 6 148 72 35 0 33.6  
## 2 1 85 66 29 0 26.6  
## 4 1 89 66 23 94 28.1  
## 5 0 137 40 35 168 43.1  
## 6 5 116 74 0 0 25.6  
## 7 3 78 50 32 88 31.0  
## DiabetesPedigreeFunction Age Outcome AgeCat  
## 1 0.627 50 1 40-50  
## 2 0.351 31 0 30-35  
## 4 0.167 21 0 21-25  
## 5 2.288 33 1 30-35  
## 6 0.201 30 0 25-30  
## 7 0.248 26 1 25-30

nrow(df)

## [1] 768

nrow(train)

## [1] 576

nrow(test)

## [1] 192

table(df$Outcome)

##   
## 0 1   
## 500 268

##Baseline accuracy

baseline <- round(500/nrow(df),2)  
baseline

## [1] 0.65

Do not select a model whose accuracy is lower than 0.65

## Fit Model using all independent variables

model = glm(Outcome~ ., data = train, family = binomial)  
summary(model)

##   
## Call:  
## glm(formula = Outcome ~ ., family = binomial, data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5680 -0.7074 -0.4311 0.7421 2.8979   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.140633 1.442614 -4.257 2.08e-05 \*\*\*  
## Pregnancies 0.043779 0.039640 1.104 0.269414   
## Glucose 0.034511 0.004255 8.111 5.00e-16 \*\*\*  
## BloodPressure -0.015237 0.006253 -2.437 0.014818 \*   
## SkinThickness 0.004535 0.007823 0.580 0.562170   
## Insulin -0.001881 0.001034 -1.819 0.068962 .   
## BMI 0.066069 0.017458 3.785 0.000154 \*\*\*  
## DiabetesPedigreeFunction 1.042778 0.353860 2.947 0.003210 \*\*   
## Age -0.048658 0.052789 -0.922 0.356659   
## AgeCat25-30 0.557992 0.418509 1.333 0.182439   
## AgeCat30-35 1.370840 0.626633 2.188 0.028697 \*   
## AgeCat35-40 1.542470 0.884172 1.745 0.081066 .   
## AgeCat40-50 2.662067 1.208167 2.203 0.027567 \*   
## AgeCat50-60 2.709033 1.748276 1.550 0.121251   
## AgeCat>60 1.850745 2.301301 0.804 0.421272   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 745.11 on 575 degrees of freedom  
## Residual deviance: 541.06 on 561 degrees of freedom  
## AIC: 571.06  
##   
## Number of Fisher Scoring iterations: 5

## Make predictions on test set

Predict <- predict(model,type = "response" , newdata = test)

## Produce Confusion matrix based on the predictions

test\_tab <- table(test$Outcome, Predict > 0.5)  
test\_tab

##   
## FALSE TRUE  
## 0 115 10  
## 1 25 42

accuracy\_test <- round(sum(diag(test\_tab))/sum(test\_tab),2)  
sprintf("Accuracy on test set is %s", accuracy\_test)

## [1] "Accuracy on test set is 0.82"