Heart Desease

Name: Austin Sampson

eMail: aws9t5@mst.edu

Course: CS 5402

Date: 02-24-2020

Concept Description:

Train a system to draw a conection between biological metrics and Chronic Heat Desiease

Data Collection:

Data has been provided from the client based off the observation of their feild agents. All data has been provided in the heart-disease.csv file ## Example Description:

Age

Scalar to represent age of the patient. zero represents the absolute lowest age. zero years old

cigsPerDay

Scalar data to represent amount of cigerates consumed a day. zero represents no cigs being used.

totChol

Scalar data to represent amout of choleseterol in the patient. zero represetns an absince of cholersterol.

sysBP

systolic blood pressure is scalar data. zero represent no blood presure(in other words death/heart attack).

diaBP

Diastolic Blood Pressure is scalar data. zero represent no blood presure(in other words death/heart attack).

BMI

body mass index is Scalar data representing expected body mass in respect to age group. zero means no body mass.

Heart Rate

Scalar Data, Zero represents the absince of a heart beat.

Blood Glucose level

Scalar Data, zero represnts an absince of Glucose in the body.

CHD

Chronic Heart Disease. This is our concept.

Data Import and Wrangling:

```
#import data
data <- read.csv("heart-disease.csv")

#impute missing values (linear regression)
imp <- mice(data, method = "norm.predict", m = 1)

##
##  iter imp variable
##   1   1  cigsPerDay  totChol  BMI  heartRate  glucose
##   2   1  cigsPerDay  totChol  BMI  heartRate  glucose
##   3   1  cigsPerDay  totChol  BMI  heartRate  glucose
##   4   1  cigsPerDay  totChol  BMI  heartRate  glucose
##   5   1  cigsPerDay  totChol  BMI  heartRate  glucose

#store data in graph form
data_imp <- complete(imp)

#Partition data set to 70% train, 30% test.
smp_size <- floor(0.70*nrow(data_imp))
set.seed(123)

train_ind <- sample(seq_len(nrow(data_imp)), size = smp_size)

#create train and test tables
train <- data_imp[train_ind, ]
test <- data_imp[-train_ind, ]
```

Mining and Analytics:

First I will begin with developing the Logistical Regression Model

```
#create Model
log_model <- glm(CHD ~., data = train, family = "binomial"(link="logit"))
#display model summary
summary(log_model)

##
## Call:
## glm(formula = CHD ~ ., family = binomial(link = "logit"), data = train)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8311  -0.5971  -0.4299  -0.3008   2.7798
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.4557408  0.6594270 -12.823  < 2e-16 ***
## age          0.0673678  0.0074243   9.074  < 2e-16 ***
## cigsPerDay   0.0304686  0.0044350   6.870 6.42e-12 ***
## totChol      0.0008665  0.0012579   0.689 0.490908
## sysBP        0.0168758  0.0037802   4.464 8.03e-06 ***
## diaBP        0.0044515  0.0070517   0.631 0.527868
## BMI          0.0003546  0.0136839   0.026 0.979327
## heartRate   -0.0072183  0.0046401  -1.556 0.119794
## glucose      0.0075786  0.0020030   3.784 0.000155 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2538.9  on 2965  degrees of freedom
## Residual deviance: 2251.4  on 2957  degrees of freedom
## AIC: 2269.4
##
## Number of Fisher Scoring iterations: 5
```

The knn operator I am using directly returns the confusion matrix rather than a model. therofer I will be covering it in the next section.

```
#K-nearst Neighbor Function K=3
nn3 <- kNN(CHD ~ .,train,test,norm=TRUE,k=3)
#confusion Matrix
table(test[,'CHD'],nn3)
```

```
##    nn3
##        0    1
##   0 1017   65
##   1  165   25
```

```
#K-nearst Neighbor Function K=1
nn2 <- kNN(CHD ~ .,train,test,norm=TRUE,k=5)
#confusion Matrix
table(test[,'CHD'],nn2)
```

```
##    nn2
##        0    1
##   0 1041   41
##   1  171   19
```

Evaluation:

logistical Regression

First I will calculate the confusion matrix for the Logistic Regression Model

```
#calculate confusion matrix
pred_log <- predict(log_model, newdata = test,type="response")

#Code Testing
test$CHD <- as.factor(test$CHD)
temp <- as.numeric(pred_log>0.5)
temp <- as.factor(temp)
#code Testing

confusionMatrix(temp, test$CHD)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1070  183
##          1   12    7
##
##                Accuracy : 0.8467
##                  95% CI : (0.8257, 0.8661)
##     No Information Rate : 0.8506
##     P-Value [Acc > NIR] : 0.6701
##
##                   Kappa : 0.0409
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.98891
##             Specificity : 0.03684
##          Pos Pred Value : 0.85395
##          Neg Pred Value : 0.36842
##              Prevalence : 0.85063
##          Detection Rate : 0.84119
##    Detection Prevalence : 0.98506
##       Balanced Accuracy : 0.51288
##
##        'Positive' Class : 0
##
```

K Nearest Neighbor

```
#K-nearst Neighbor Function K=3
nn3 <- kNN(CHD ~ .,train,test,norm=TRUE,k=3)
```

```
#confusion Matrix
table(test[,'CHD'],nn3)
```

```
##    nn3
##        0    1
##   0 1017   65
##   1  165   25
```

Error Rate = (65+165)/(1017+65+165+25)=0.1808

Accuracy = (1017+25)/(1017+65+165+25)=0.8192

Precission= (1017)/(1017+65)=0.9621

Recall=(1017)/(1017+165)=0.9399

F-measure=(2*0.9399*0.9621)/(0.8589+0.9621)=0.9075

Based off the results of the confusion matrices for the two model I would present the Logistical Regression model to the customer. I would do this because the logistical regressional model presents a slightly better accuracy but more importantly the model produces significantly less false negatives. Due to the nature of what we are predicting being corolated to the risk of heart attack or stroke we should prioritize minimizing false negatives because they present an increased risk of death. False positives can be identified with further medical tests.

Referinces:

https://cran.r-project.org/web/packages/mice/mice.pdf
https://stats.stackexchange.com/questions/100841/imputation-by-regression-in-r
https://stackoverflow.com/questions/17200114/how-to-split-data-into-training-testing-sets-using-sample-function

https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/kNN
https://stats.idre.ucla.edu/r/dae/logit-regression/ https://stats.idre.ucla.edu/r/dae/logit-regression/ https://intellipaat.com/community/1546/error-in-confusion-matrix-the-data-and-reference-factors-must-have-the-same-number-of-levels