Classification

Rikita Patangay

2022-09-25

Intro to Classification

In classification the target will be a binary output that we classify into one class or the other. There are techniques that allow classification into more than two classes.

Strengths: - Separates classes well if they are capable of being seprerated - Computationally inexpensive - Nice probabilistic output

Weaknesses: - It is prone to underfitting.

Read in Dataset: International Matches (FIFA World Cup)

Using read.csv to read in the file and put the data set in the variable (df).

```
df <- read.csv("international_matches.csv")
str(df)
```

```
## 'data.frame':    23921 obs. of  25 variables:
##  $ date                          : chr  "1993-08-08" "1993-08-08" "1993-08-08" "1993-08-08" ...
##  $ home_team                     : chr  "Bolivia" "Brazil" "Ecuador" "Guinea" ...
##  $ away_team                     : chr  "Uruguay" "Mexico" "Venezuela" "Sierra Leone" ...
##  $ home_team_continent           : chr  "South America" "South America" "South America" "Africa" ...
##  $ away_team_continent           : chr  "South America" "North America" "South America" "Africa" ...
##  $ home_team_fifa_rank           : int  59 8 35 65 67 70 50 65 111 4 ...
##  $ away_team_fifa_rank           : int  22 14 94 86 5 19 102 86 9 3 ...
##  $ home_team_total_fifa_points   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ away_team_total_fifa_points   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ home_team_score               : int  3 1 5 1 1 0 2 4 0 1 ...
##  $ away_team_score               : int  1 1 0 0 3 1 0 0 7 2 ...
##  $ tournament                    : chr  "FIFA World Cup qualification" "Friendly" "FIFA World Cup qual
##  $ city                          : chr  "La Paz" "Maceió" "Quito" "Conakry" ...
##  $ country                       : chr  "Bolivia" "Brazil" "Ecuador" "Guinea" ...
##  $ neutral_location              : chr  "False" "False" "False" "False" ...
##  $ shoot_out                     : chr  "No" "No" "No" "No" ...
##  $ home_team_result              : chr  "Win" "Draw" "Win" "Win" ...
##  $ home_team_goalkeeper_score    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ away_team_goalkeeper_score    : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ home_team_mean_defense_score : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ home_team_mean_offense_score : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ home_team_mean_midfield_score: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ away_team_mean_defense_score : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ away_team_mean_offense_score : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ away_team_mean_midfield_score: num  NA NA NA NA NA NA NA NA NA NA ...
```

a.) Split data 80/20

Here I am dividing the data in training and test sets. This works by randomly sampling the data using the sample() function. This is an 80/20 split.

```
set.seed(1234)
i <- sample(1:nrow(df), .80*nrow(df), replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

b.) 5 Functions

These are 5 pretty simple functions used for data exploration.

head() - gives some of the start of the data sample

tail() - gives some of the end of the data sample

nrow() - gives number of rows in the sample

ncol() - gives number of columns in the sample

summary() - gives a brief summary of values associated with the data.

```
head(train)
```

```
##            date home_team     away_team home_team_continent
## 7452  2003-10-22   Myanmar          Iraq                Asia
## 8016  2004-06-06      Peru     Venezuela       South America
## 7162  2003-06-29    France      Cameroon              Europe
## 8086  2004-06-20     Egypt Côte d'Ivoire              Africa
## 23653 2022-05-31  Thailand       Bahrain                Asia
## 9196  2005-10-08    Latvia         Japan              Europe
##       away_team_continent home_team_fifa_rank away_team_fifa_rank
## 7452                 Asia                 141                  52
## 8016        South America                  81                  49
## 7162               Africa                   2                  16
## 8086               Africa                  29                  69
## 23653                Asia                 111                  89
## 9196                 Asia                  63                  16
##       home_team_total_fifa_points away_team_total_fifa_points home_team_score
## 7452                            0                           0               1
## 8016                            0                           0               0
## 7162                            0                           0               1
## 8086                            0                           0               1
## 23653                        1167                        1262               1
## 9196                            0                           0               2
##       away_team_score                  tournament         city  country
## 7452                3  AFC Asian Cup qualification        Riffa  Bahrain
## 8016                0 FIFA World Cup qualification         Lima     Peru
## 7162                0           Confederations Cup  Saint-Denis   France
## 8086                2 FIFA World Cup qualification   Alexandria    Egypt
## 23653               2                    Friendly Pathum Thani Thailand
## 9196                2                    Friendly         Riga   Latvia
##       neutral_location shoot_out home_team_result home_team_goalkeeper_score
```

```
## 7452              True        No             Lose                          NA
## 8016             False        No             Draw                          NA
## 7162             False        No              Win                          NA
## 8086             False        No             Lose                          NA
## 23653            False        No             Lose                          66
## 9196             False        No             Draw                          NA
##       away_team_goalkeeper_score home_team_mean_defense_score
## 7452                           NA                           NA
## 8016                           NA                           NA
## 7162                           NA                           NA
## 8086                           NA                           NA
## 23653                          NA                           NA
## 9196                           NA                           NA
##       home_team_mean_offense_score home_team_mean_midfield_score
## 7452                            NA                           NA
## 8016                            NA                           NA
## 7162                            NA                           NA
## 8086                            NA                           NA
## 23653                           NA                           NA
## 9196                            NA                           NA
##       away_team_mean_defense_score away_team_mean_offense_score
## 7452                            NA                           NA
## 8016                            NA                           NA
## 7162                            NA                           NA
## 8086                            NA                           NA
## 23653                           NA                           NA
## 9196                            NA                         70.3
##       away_team_mean_midfield_score
## 7452                             NA
## 8016                             NA
## 7162                             NA
## 8086                             NA
## 23653                            NA
## 9196                           74.8
```

```
tail(train)
```

```
##             date            home_team  away_team home_team_continent
## 2963  1998-03-29             Colombia   Paraguay       South America
## 15548 2012-11-20             Malaysia Bangladesh                Asia
## 13297 2010-08-29             Ethiopia       Chad              Africa
## 5446  2001-04-23         Turkmenistan     Jordan                Asia
## 17121 2014-10-10 United Arab Emirates  Australia                Asia
## 17547 2015-03-29               France    Denmark              Europe
##       away_team_continent home_team_fifa_rank away_team_fifa_rank
## 2963        South America                 17                  30
## 15548                Asia                163                 171
## 13297              Africa                144                 124
## 5446                 Asia                131                 109
## 17121             Oceania                 73                  84
## 17547              Europe                  8                  28
##       home_team_total_fifa_points away_team_total_fifa_points home_team_score
## 2963                            0                           0               1
## 15548                         126                         113               1
```

```
## 13297                          0                          0              1
## 5446                           0                          0              2
## 17121                        429                        390              0
## 17547                       1179                        863              2
##       away_team_score                  tournament          city
## 2963                1                    Friendly     New Haven
## 15548               1                    Friendly  Kuala Lumpur
## 13297               0                    Friendly   Addis Ababa
## 5446                0 FIFA World Cup qualification      Tashkent
## 17121               0                    Friendly     Abu Dhabi
## 17547               0                    Friendly Saint-Étienne
##                    country neutral_location shoot_out home_team_result
## 2963                   USA             True        No             Draw
## 15548              Malaysia            False        No             Draw
## 13297              Ethiopia            False        No              Win
## 5446             Uzbekistan             True        No              Win
## 17121 United Arab Emirates            False        No             Draw
## 17547                France            False        No              Win
##       home_team_goalkeeper_score away_team_goalkeeper_score
## 2963                          NA                         NA
## 15548                         NA                         NA
## 13297                         NA                         NA
## 5446                          NA                         NA
## 17121                         NA                         76
## 17547                         85                         77
##       home_team_mean_defense_score home_team_mean_offense_score
## 2963                            NA                           NA
## 15548                           NA                           NA
## 13297                           NA                           NA
## 5446                            NA                           NA
## 17121                           NA                           NA
## 17547                           81                         81.3
##       home_team_mean_midfield_score away_team_mean_defense_score
## 2963                             NA                           NA
## 15548                            NA                           NA
## 13297                            NA                           NA
## 5446                             NA                           NA
## 17121                            NA                         70.0
## 17547                          84.5                         77.2
##       away_team_mean_offense_score away_team_mean_midfield_score
## 2963                            NA                           NA
## 15548                           NA                           NA
## 13297                           NA                           NA
## 5446                            NA                           NA
## 17121                         71.7                         71.2
## 17547                         75.0                         77.0
```

```
nrow(train)
```

```
## [1] 19136
```

```
ncol(train)
```

```
## [1] 25
```

```
summary(train)
```

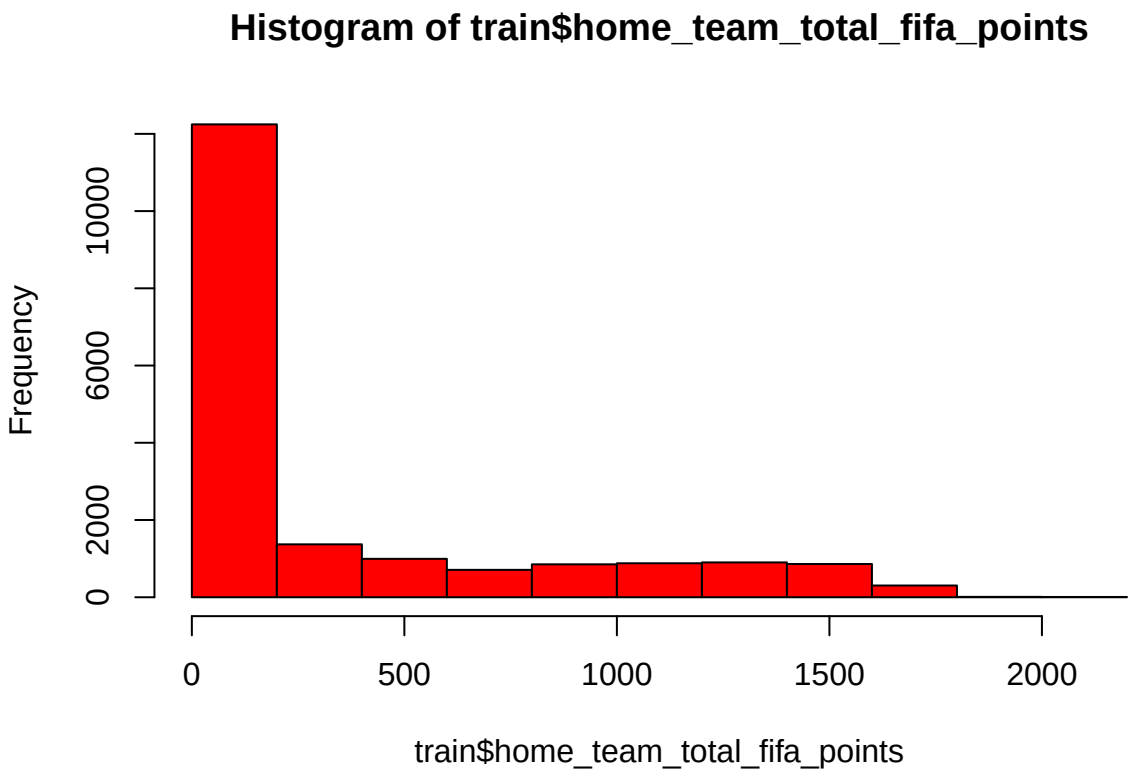
```
##      date            home_team          away_team         home_team_continent
##  Length:19136       Length:19136       Length:19136       Length:19136
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  away_team_continent home_team_fifa_rank away_team_fifa_rank
##  Length:19136        Min.   :  1.0       Min.   :  1.00
##  Class :character    1st Qu.: 33.0       1st Qu.: 36.00
##  Mode  :character    Median : 71.0       Median : 73.00
##                      Mean   : 77.5       Mean   : 80.68
##                      3rd Qu.:115.0       3rd Qu.:119.00
##                      Max.   :211.0       Max.   :211.00
##
##  home_team_total_fifa_points away_team_total_fifa_points home_team_score
##  Min.   :   0.0              Min.   :   0.0              Min.   : 0.000
##  1st Qu.:   0.0              1st Qu.:   0.0              1st Qu.: 0.000
##  Median :   0.0              Median :   0.0              Median : 1.000
##  Mean   : 326.2              Mean   : 317.5              Mean   : 1.611
##  3rd Qu.: 555.0              3rd Qu.: 525.2              3rd Qu.: 2.000
##  Max.   :2164.0              Max.   :2124.0              Max.   :22.000
##
##  away_team_score   tournament           city             country
##  Min.   : 0.000   Length:19136       Length:19136       Length:19136
##  1st Qu.: 0.000   Class :character   Class :character   Class :character
##  Median : 1.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 1.062
##  3rd Qu.: 2.000
##  Max.   :21.000
##
##  neutral_location    shoot_out         home_team_result
##  Length:19136       Length:19136       Length:19136
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  home_team_goalkeeper_score away_team_goalkeeper_score
##  Min.   :47.00              Min.   :47.00
##  1st Qu.:70.00              1st Qu.:69.00
##  Median :75.00              Median :74.00
##  Mean   :75.06              Mean   :74.26
##  3rd Qu.:81.00              3rd Qu.:80.00
##  Max.   :97.00              Max.   :97.00
##  NA's   :12366              NA's   :12603
##  home_team_mean_defense_score home_team_mean_offense_score
##  Min.   :52.80                Min.   :53.3
##  1st Qu.:71.00                1st Qu.:71.7
```

```
##  Median :75.20                Median :75.7
##  Mean   :74.97                Mean   :75.9
##  3rd Qu.:78.80                3rd Qu.:80.0
##  Max.   :91.80                Max.   :93.0
##  NA's   :12836                NA's   :12267
##  home_team_mean_midfield_score away_team_mean_defense_score
##  Min.   :55.50                 Min.   :52.80
##  1st Qu.:72.50                 1st Qu.:70.50
##  Median :76.20                 Median :74.50
##  Mean   :75.93                 Mean   :74.48
##  3rd Qu.:79.50                 3rd Qu.:78.20
##  Max.   :93.20                 Max.   :91.80
##  NA's   :12523                 NA's   :13032
##  away_team_mean_offense_score away_team_mean_midfield_score
##  Min.   :53.30                Min.   :54.2
##  1st Qu.:71.30                1st Qu.:71.8
##  Median :75.30                Median :75.8
##  Mean   :75.45                Mean   :75.3
##  3rd Qu.:79.70                3rd Qu.:79.0
##  Max.   :93.00                Max.   :93.2
##  NA's   :12439                NA's   :12688
```

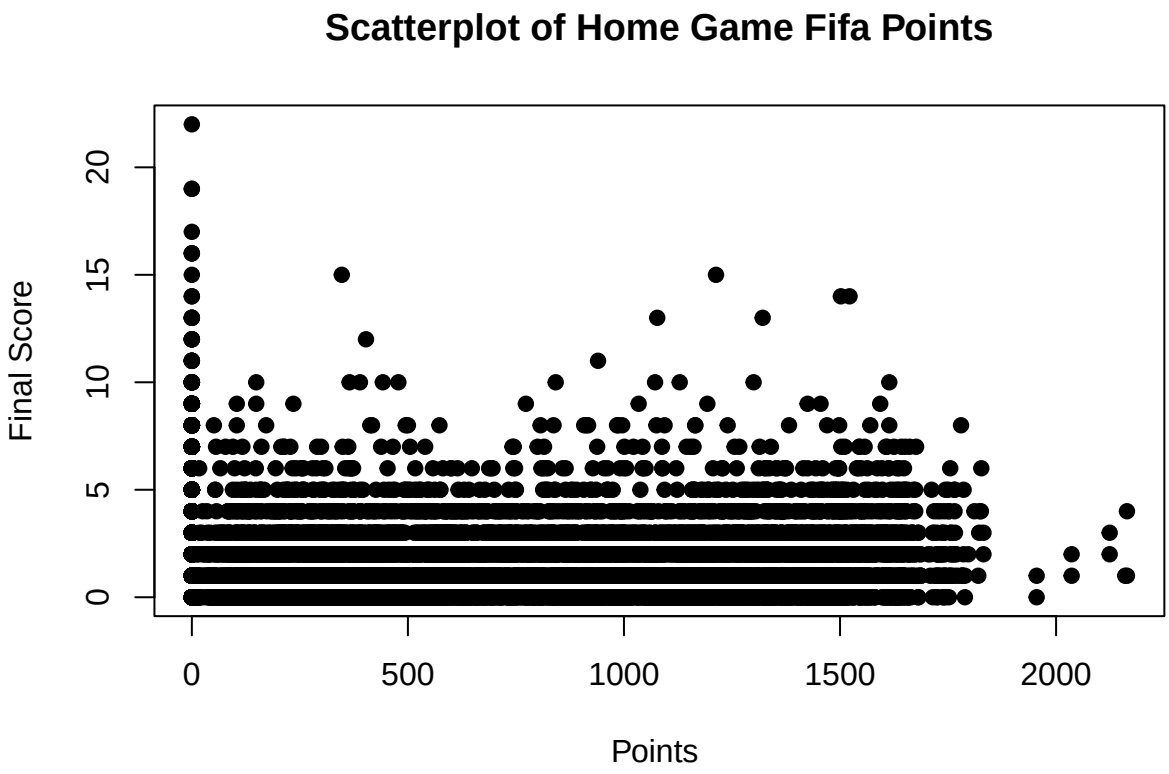
c.) 2 Informative Graphs

These are two informative graphs. The first one is a histogram and the second is scatterplot.

```
hist(train$home_team_total_fifa_points, breaks=12, col="red")
```




```
attach(train)
plot(home_team_total_fifa_points, home_team_score, main="Scatterplot of Home Game Fifa Points",
   xlab="Points ", ylab="Final Score", pch=19)
```



d.) Build a Logistic Regression Model

Here we are building a logistic regression model. What we are doing in the model is using the predictor which would be the Rel.Hum. Which is the relative humidity, and our target (y) is Visibility.

e.) Naïve Bayes Model

Based of the Naïve Bayes Model I can say that it is showing me the data of Draws, Losses and Win for randomized teams. (I would copy and paste the data but there is too much and it would get unformatted.)

f.) Predict and Evaluate

I predict that the values of the home team score and final result score are very close together. Evaluating the data shows that my prediction is accurate.

g.) Strengths and Weaknesses of Naïve Bayes Model and Logistic Regression

Strengths: - Works well with small data sets - Easy to implement - Easy to interpret - Handles high dimensions well

Weaknesses: - May be outperformed by other classifiers for larger data sets - Guess values in the test set that are not in the training data - If the predictors are not independent, the naive assumption that they are may limit the performance of the algorithm

h.) Classification metrics

Accuracy, Recall, Precision, and F1- Score are the metrics that were measured.