# Report

SOFIN WADHWANIYA

28/10/2020

## Logistic regression

### *Report of your Regression Model*

#### About Logistic Regression

Logistic Regression, also known as Logit Regression or Logit Model, is a mathematical model used in statistics to estimate (guess) the probability of an event occurring having been given some previous data. Logistic Regression works with binary data, where either the event happens (1) or the event does not happen (0).

Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Here, the response variables can be categorical or continuous, as the model does not strictly require continuous data. To predict group membership, logistic regression uses the log odds ratio rather than probabilities and an iterative maximum likelihood method rather than a least squares to fit the final model. This means the researcher has more freedom when using logistic regression and the method may be more appropriate for non-normally distributed data or when the samples have unequal co-variance matrices.

Thus, Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems.

#### Overview of Data

```
##   Survived Pclass Age    Fare
## 1        0      3  22  7.2500
## 2        1      1  38 71.2833
## 3        1      3  26  7.9250
## 4        1      1  35 53.1000
## 5        0      3  35  8.0500
## 7        0      1  54 51.8625
```

#### Summary of Data

```
## Warning in png(png_loc <- tempfile(fileext = ".png"), width = 150 *
## graph.magnif, : unable to open connection to X11 display ''
```

```
## Warning in png(png_loc <- tempfile(fileext = ".png"), width = 150 *
## graph.magnif, : unable to open connection to X11 display ''
```

```
## Warning in png(png_loc <- tempfile(fileext = ".png"), width = 150 *
## graph.magnif, : unable to open connection to X11 display ''

## Warning in png(png_loc <- tempfile(fileext = ".png"), width = 150 *
## graph.magnif, : unable to open connection to X11 display ''
```

```
## Data Frame Summary
## dtf2()
## Dimensions: 637 x 4
## Duplicates: 43
##
## ----------------------------------------------------------------------------------------------------
## No   Variable    Stats / Values             Freqs (% of Valid)    Graph            Valid    Missing
## ---- ----------- -------------------------- --------------------- ---------------- -------- ---------
## 1    Survived    Min  : 0                   0 : 378 (59.3%)       IIIIIIIIIII      637      0
##      [integer]   Mean : 0.4                 1 : 259 (40.7%)       IIIIIIII         (100%)   (0%)
##                  Max  : 1
##
## 2    Pclass      Mean (sd) : 2.2 (0.8)      1 : 168 (26.4%)       IIIII            637      0
##      [integer]   min < med < max:           2 : 155 (24.3%)       IIII             (100%)   (0%)
##                  1 < 2 < 3                  3 : 314 (49.3%)       IIIIIIIII
##                  IQR (CV) : 2 (0.4)
##
## 3    Age         Mean (sd) : 29.9 (14.5)    85 distinct values        :            637      0
##      [numeric]   min < med < max:                                     :            (100%)   (0%)
##                  0.7 < 28 < 80                                      . : :
##                  IQR (CV) : 17 (0.5)                                . : : : :
##                                                                     : : : : : : : .
##
## 4    Fare        Mean (sd) : 36 (55.1)      207 distinct values   :                637      0
##      [numeric]   min < med < max:                                 :                (100%)   (0%)
##                  0 < 15.9 < 512.3                                :
##                  IQR (CV) : 26.3 (1.5)                           :
##                                                                  : : .
## ----------------------------------------------------------------------------------------------------
```

### Structure of Data

```
## 'data.frame':    637 obs. of  4 variables:
##  $ Survived: int  0 1 1 1 0 0 0 1 1 1 ...
##  $ Pclass  : int  3 1 3 1 3 1 3 3 2 3 ...
##  $ Age     : num  22 38 26 35 35 54 2 27 14 4 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  - attr(*, "na.action")= 'omit' Named int [1:163] 6 18 20 27 29 30 32 33 37 43 ...
##   ..- attr(*, "names")= chr [1:163] "6" "18" "20" "27" ...
```

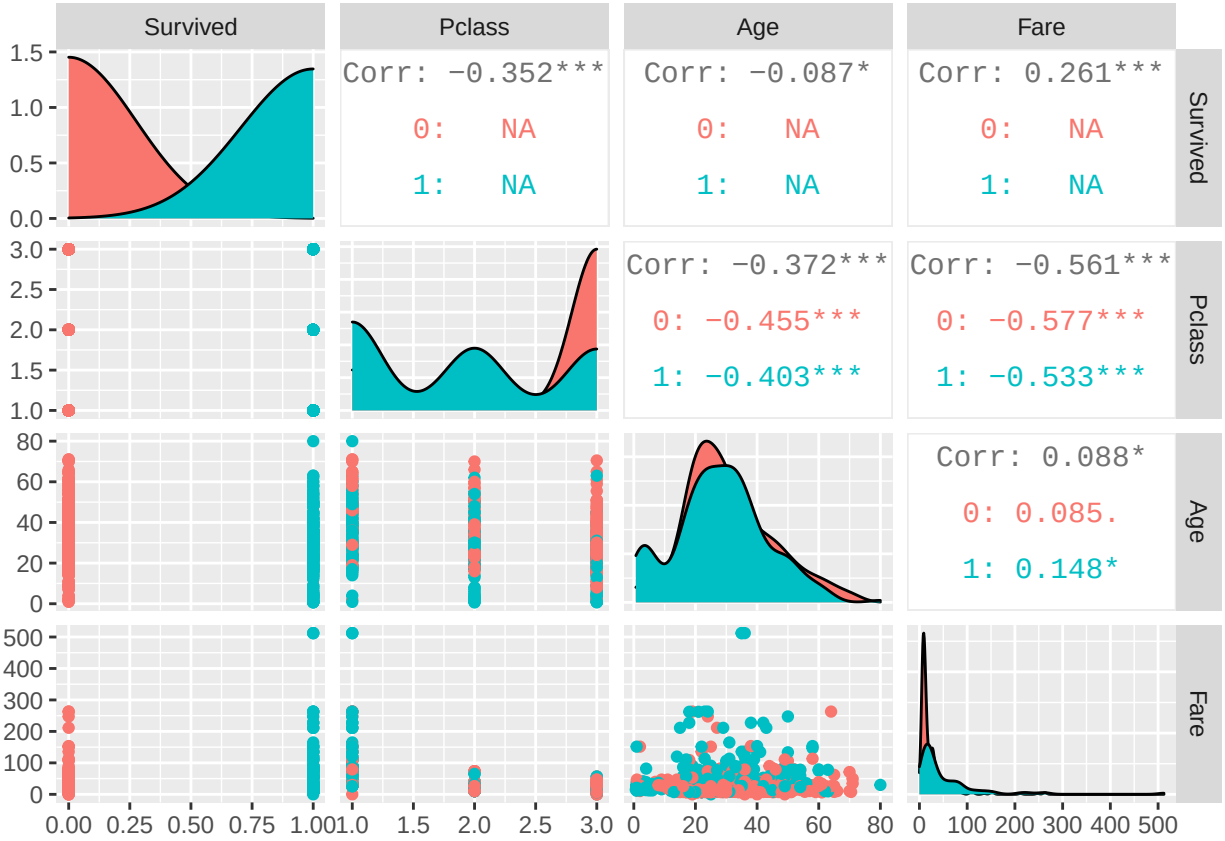
### Descriptive Statistics

```
## Descriptive Statistics
## dtf2()
## N: 637
##
##                       Age     Fare   Pclass   Survived
```

```
## ----------------- -------- -------- -------- ----------
##              Mean    29.87    35.95     2.23       0.41
##           Std.Dev    14.54    55.06     0.84       0.49
##               Min     0.67     0.00     1.00       0.00
##                Q1    21.00     8.05     1.00       0.00
##            Median    28.00    15.90     2.00       0.00
##                Q3    38.00    34.38     3.00       1.00
##               Max    80.00   512.33     3.00       1.00
##               MAD    12.60    12.23     1.48       0.00
##               IQR    17.00    26.32     2.00       1.00
##                CV     0.49     1.53     0.38       1.21
##          Skewness     0.39     4.52    -0.45       0.38
##       SE.Skewness     0.10     0.10     0.10       0.10
##          Kurtosis     0.12    28.66    -1.44      -1.86
##           N.Valid   637.00   637.00   637.00     637.00
##         Pct.Valid   100.00   100.00   100.00     100.00
```

**Pair Plot as Visualisation**

```
## Warning in cor(x, y): the standard deviation is zero

## Warning in cor(x, y): the standard deviation is zero

## Warning in cor(x, y): the standard deviation is zero

## Warning in cor(x, y): the standard deviation is zero

## Warning in cor(x, y): the standard deviation is zero

## Warning in cor(x, y): the standard deviation is zero
```




### Regression

### Model Summary

```
##
## Call:
## glm(formula = dtf2()[, input$columns] ~ ., family = binomial(link = "logit"),
##     data = dtf3())
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2323  -0.8628  -0.6234   1.0023   2.3799
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.294404   0.517357   6.368 1.92e-10 ***
## Pclass      -1.144886   0.150193  -7.623 2.48e-14 ***
## Age         -0.042102   0.007183  -5.862 4.59e-09 ***
## Fare         0.002242   0.002233   1.004    0.315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 860.71  on 636  degrees of freedom
## Residual deviance: 738.30  on 633  degrees of freedom
```

```
## AIC: 746.3
##
## Number of Fisher Scoring iterations: 4
```

### Coefficients

```
##
## Call:  glm(formula = dtf2()[, input$columns] ~ ., family = binomial(link = "logit"),
##     data = dtf3())
##
## Coefficients:
## (Intercept)       Pclass          Age         Fare
##    3.294404    -1.144886    -0.042102     0.002242
##
## Degrees of Freedom: 636 Total (i.e. Null);  633 Residual
## Null Deviance:       860.7
## Residual Deviance: 738.3     AIC: 746.3
```

### ANOVA Table

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: dtf2()[, input$columns]
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     636     860.71
## Pclass  1   79.995       635     780.71 < 2.2e-16 ***
## Age     1   41.301       634     739.41 1.305e-10 ***
## Fare    1    1.113       633     738.30    0.2915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

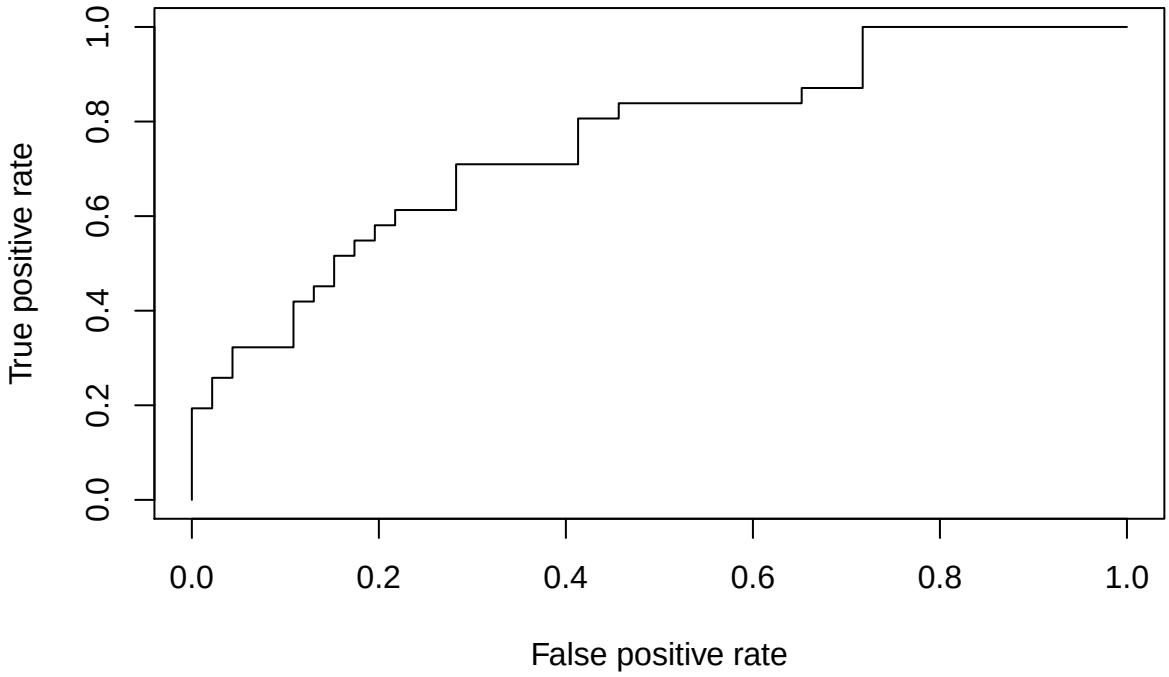
### Model Assessment on your Test data

### Confusion Matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 39 16
##          1  7 15
##
##                Accuracy : 0.7013
##                  95% CI : (0.5862, 0.8003)
##     No Information Rate : 0.5974
##     P-Value [Acc > NIR] : 0.03894
##
```

```
##                   Kappa : 0.3482
##
##  Mcnemar's Test P-Value : 0.09529
##
##             Sensitivity : 0.8478
##             Specificity : 0.4839
##          Pos Pred Value : 0.7091
##          Neg Pred Value : 0.6818
##              Prevalence : 0.5974
##          Detection Rate : 0.5065
##    Detection Prevalence : 0.7143
##       Balanced Accuracy : 0.6658
##
##        'Positive' Class : 0
##
```

**Performance Instance Plot (True vs False Positive Rate)**




**Contact Details**

Name: Sofin Wadhwaniya"

Email address: sofinwadhwaniya@gmail.com