“A project on Logistic Regression Predictive model”

Contents

SECTION1: Problem Definition

Build a model to predict if a pregnant woman is diabetic or not. Apply Logistic Regression technique to predict the binary Outcome.

SECTION2: Brief description of the data

The in this project has been taken from the website www.kaggle.com .

The data was collected and made available by "National Institute of Diabetes and Digestive and Kidney Diseases" as part of the Pima Indians Diabetes Database. All patients for which data is collected belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above.

The collected data is containing record for 768 pregnant women and 8 different variables and Outcome (0,1) as dependent variable.

Actual Data:

diabetes2.csv

SECTION 3: Data Visualization

summary(data_diabetes)

(data_diabetes$Pregnancies)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   3.000   3.845   6.000  17.000
```

summary(data_diabetes$Glucose) #Doesn't make sense to have Glucose 0 for any living human being. It makes sense to replace ZERO's with mean value

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    99.0   117.0   120.9   140.2   199.0
```

summary(data_diabetes$BloodPressure) #Doesn't make sense to have BP 0 for any living human being. I t makes sense to replace ZERO's with mean value

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   62.00   72.00   69.11   80.00  122.00
```

summary(data_diabetes$SkinThickness)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00   23.00   20.54   32.00   99.00
```

summary(data_diabetes$Insulin)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0     0.0    30.5    79.8   127.2   846.0
```

summary(data_diabetes$BMI) #Doesn't make sense to have BMI 0 for any living human being. It makes sense to replace ZERO's with mean value

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   27.30   32.00   31.99   36.60   67.10
```

summary(data_diabetes$DiabetesPedigreeFunction)

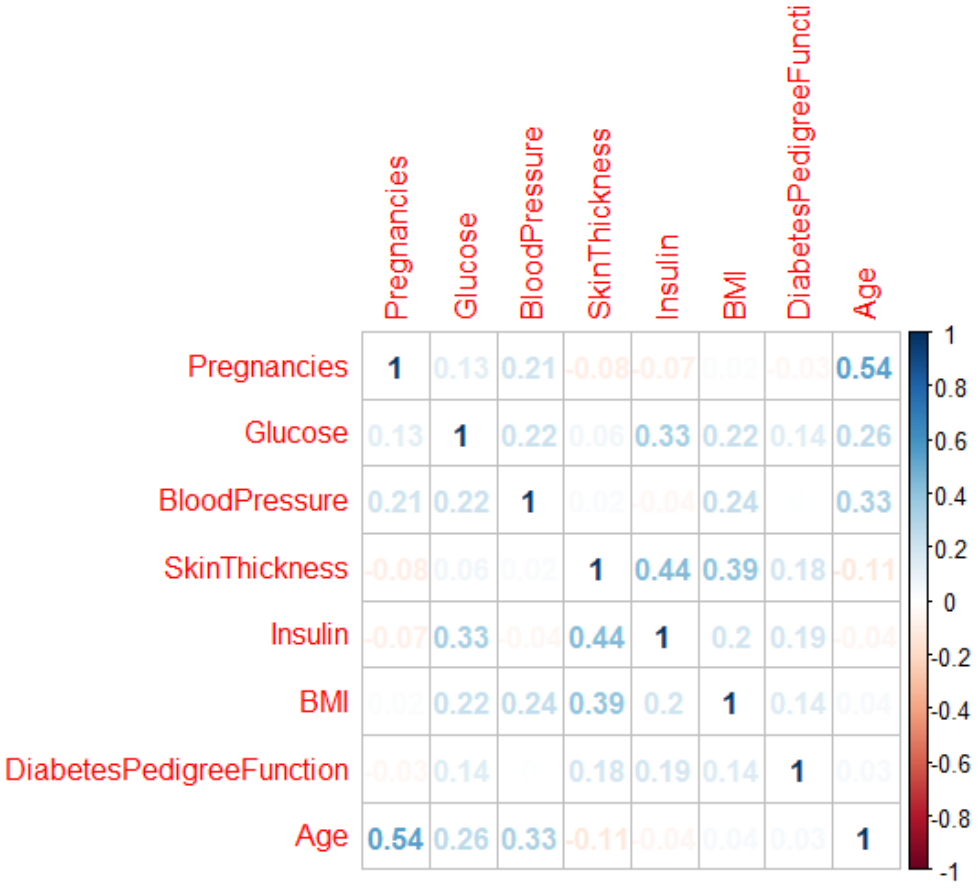
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0780  0.2437  0.3725  0.4719  0.6262  2.4200
```

```
summary(data_diabetes$Age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  21.00   24.00   29.00   33.24   41.00   81.00
```

Check for Multi-collinearity

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Pregnancies	1.00000000	0.12945867	0.14128198	-0.08167177	-0.07353461	0.01768309	-0.03352267	0.54434123
Glucose	0.12945867	1.00000000	0.15258959	0.05732789	0.33135711	0.22107107	0.13733730	0.26351432
BloodPressure	0.14128198	0.15258959	1.00000000	0.20737054	0.08893338	0.28180529	0.04126495	0.23952795
SkinThickness	-0.08167177	0.05732789	0.20737054	1.00000000	0.43678257	0.39257320	0.18392757	-0.11397026
Insulin	-0.07353461	0.33135711	0.08893338	0.43678257	1.00000000	0.19785906	0.18507093	-0.04216295
BMI	0.01768309	0.22107107	0.28180529	0.39257320	0.19785906	1.00000000	0.14064695	0.03624187
DiabetesPedigreeFunction	-0.03352267	0.13733730	0.04126495	0.18392757	0.18507093	0.14064695	1.00000000	0.03356131
Age	0.54434123	0.26351432	0.23952795	-0.11397026	-0.04216295	0.03624187	0.03356131	1.00000000



SECTION 4: Regression Refinements:

Model 1: #Regression model with all variables.

```
glm(formula = dc.train$Outcome ~ dc.train$Pregnancies + dc.train$Glucose +
    dc.train$BloodPressure + dc.train$SkinThickness + dc.train$Insulin +
    dc.train$BMI + dc.train$DiabetesPedigreeFunction + dc.train$Age,
```

```
    family = binomial(link  = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3987  -0.7337  -0.4270   0.7077   2.8948

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -8.562672   0.946029  -9.051  < 2e-16 ***
dc.train$Pregnancies           0.159363   0.039774   4.007 6.16e-05 ***
dc.train$Glucose               0.036675   0.004578   8.011 1.13e-15 ***
dc.train$BloodPressure        -0.012091   0.009853  -1.227   0.2198
dc.train$SkinThickness         0.004994   0.008333   0.599   0.5490
dc.train$Insulin              -0.002282   0.001201  -1.901   0.0573 .
dc.train$BMI                   0.095691   0.018255   5.242 1.59e-07 ***
dc.train$DiabetesPedigreeFunction 0.631605   0.354898   1.780   0.0751 .
dc.train$Age                   0.006943   0.011390   0.610   0.5421
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 696.28  on 537  degrees of freedom
Residual deviance: 510.71  on 529  degrees of freedom
AIC: 528.71
Number of Fisher Scoring iterations: 5
```

Model 2: With less number of variables:

#We can ignore below variables as p >>> .05

#BloodPressure #SkinThickness #Insulin #Age # DiabetesPedigreeFunction

New model after ignoring above variables -

```
glm(formula = dc.train$Outcome ~ dc.train$Pregnancies + dc.train$Glucose +
    dc.train$BMI, family = binomial(link  = "logit"))


Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2854  -0.7445  -0.4226   0.7140   2.8233
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-8.564176	0.797769	-10.735	< 2e-16	***
dc.train$Pregnancies	0.167073	0.033221	5.029	4.93e-07	***
dc.train$Glucose	0.033890	0.003971	8.534	< 2e-16	***
dc.train$BMI	0.092178	0.016547	5.571	2.54e-08	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 696.28 on 537 degrees of freedom

Residual deviance: 518.04 on 534 degrees of freedom

AIC: 526.04

Number of Fisher Scoring iterations: 5

SECTION5: Conclusion

1. By Comparing AIC of 2 values: AIC: 528.71 model1 and AIC: 526.04of model2
    a. Model 2 is better, lower AIC it better the model.
2. Model accuracy: **76% (for Training data)**

	Predicted values		
Actual		0	1
Values	0	310	40
	1	83	105

Accuracy: 415/538 = 0.7713 = 77.13%

3. Prediction for test data: **75.65% (for Test Data)** which is same as training data.

	Predicted for test data		
		0	1
Actual	0	126	24
values	1	32	48

4. Vif (reg.mod1)

dc.train$Pregnancies	dc.train$Glucose	dc.train$BMI
1.036933	1.005040	1.032172

SECTION6: R File (code)

DiabeteseDetection R.R