YAML

Fumilayo Moustapha ANA 515 Assignment 2

2022

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

GOAL: The aim of this R project is to build a classifier that can detect credit card fraudulent transactions. We will use a variety of machine learning algorithms that will be able to discern fraudulent from non-fraudulent one. By the end of this machine learning project, you will learn how to implement machine learning algorithms to perform classification.

DATA RETRIEVE AND VIEW OF DATA dataset retrieve from data flair project website and data set was reteive from the project

4 description of data

5 DATA PREPARATION For our data we want to run the logistic regression model to predict the fraudelent transaction so we divide our data into two subset based on spilt ratio 0.8 and name the two dataset as train_data and test_ data and we describe the dimension of the both data which are equal so that we can run and perform the logistics regression on one dataset easily

```
library(caTools)
set.seed(123)
data_sample = sample.split(NewData$Class,SplitRatio=0.80)
train_data = subset(NewData,data_sample==TRUE)
test_data = subset(NewData,data_sample==FALSE)
dim(train_data)
```

```
## [1] 227846     30
```

```
dim(test_data)
```

```
## [1] 56961    30
```

6 WE RUN THE LOGISTICS REGRESSION on our data because we want to acheivew that which transaction looks fraudelent in the batch of 6 month period and we apply the binary logistics model on our data.

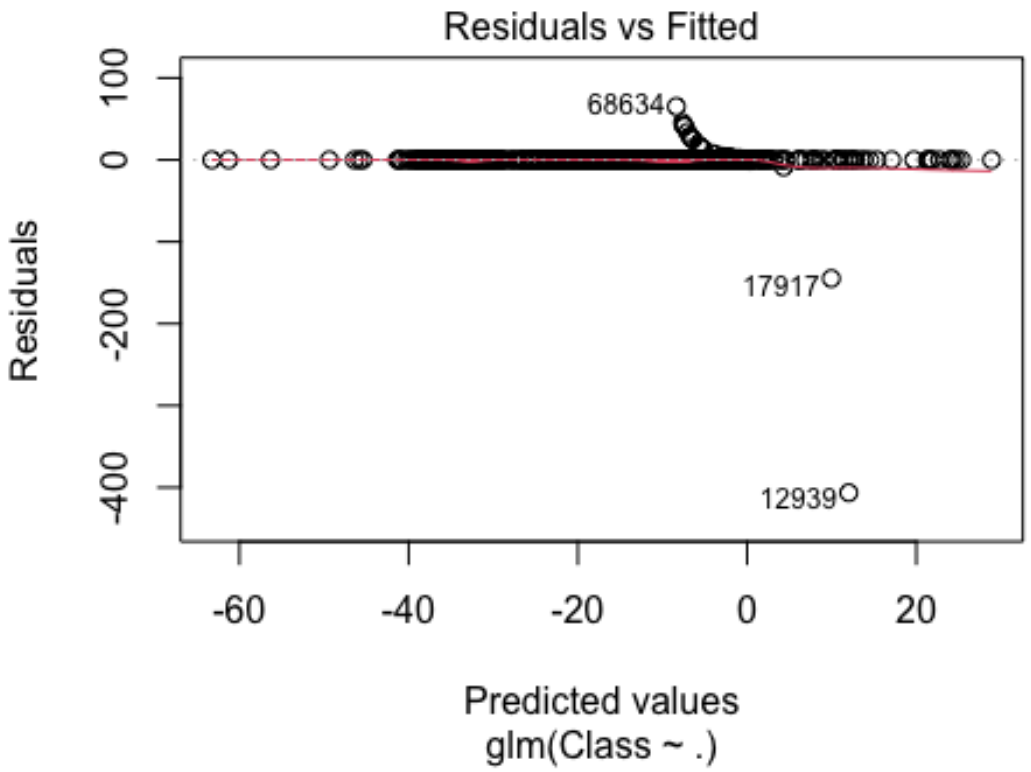
```
Logistic_Model=glm(Class~.,test_data,family=binomial())

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

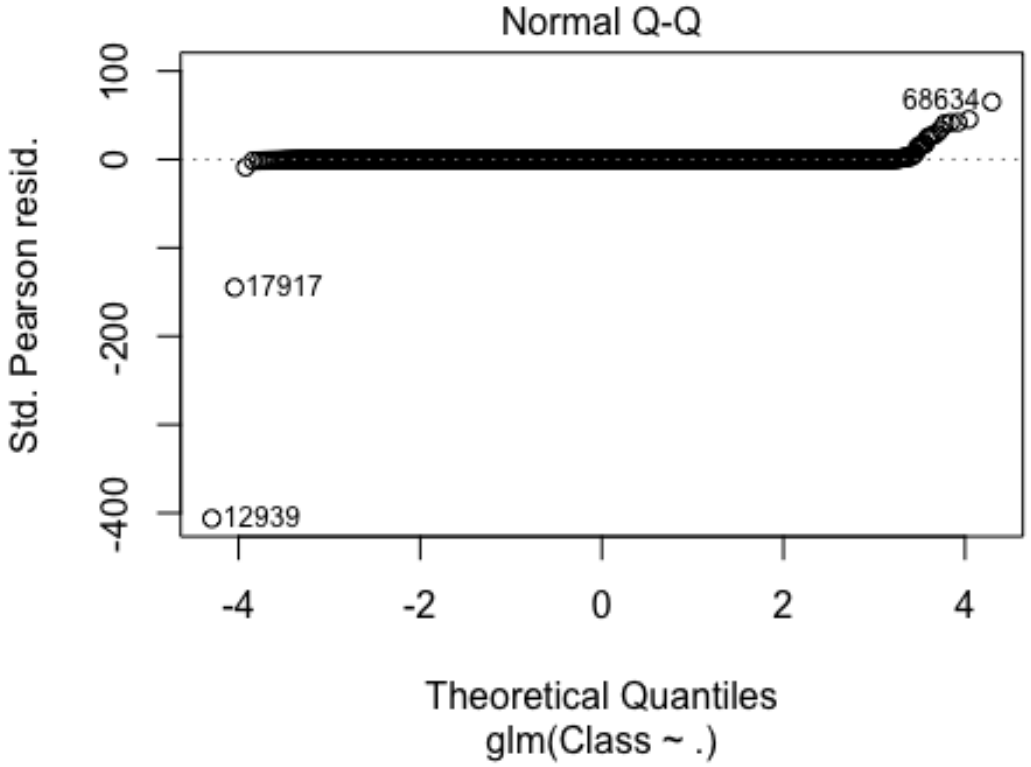
7 SUMMARY we first observe the deviance residuals of the class of the trasaction that will access the fitting of the model so we say that 5 number summary of the model as the deviation of residual is minimum = -4.9019 and first quartile is -0.0254 and median is -0.0156 and the max is 4.0877 so the range is about 9.7096 which means the data have less deviation from the mean.we see the coefficient of different algorithm which is used to predict the fraudlent activity so we can say that all coefficient are insignificant while V8 observe some fraudlent trasaction which look significant in the model.

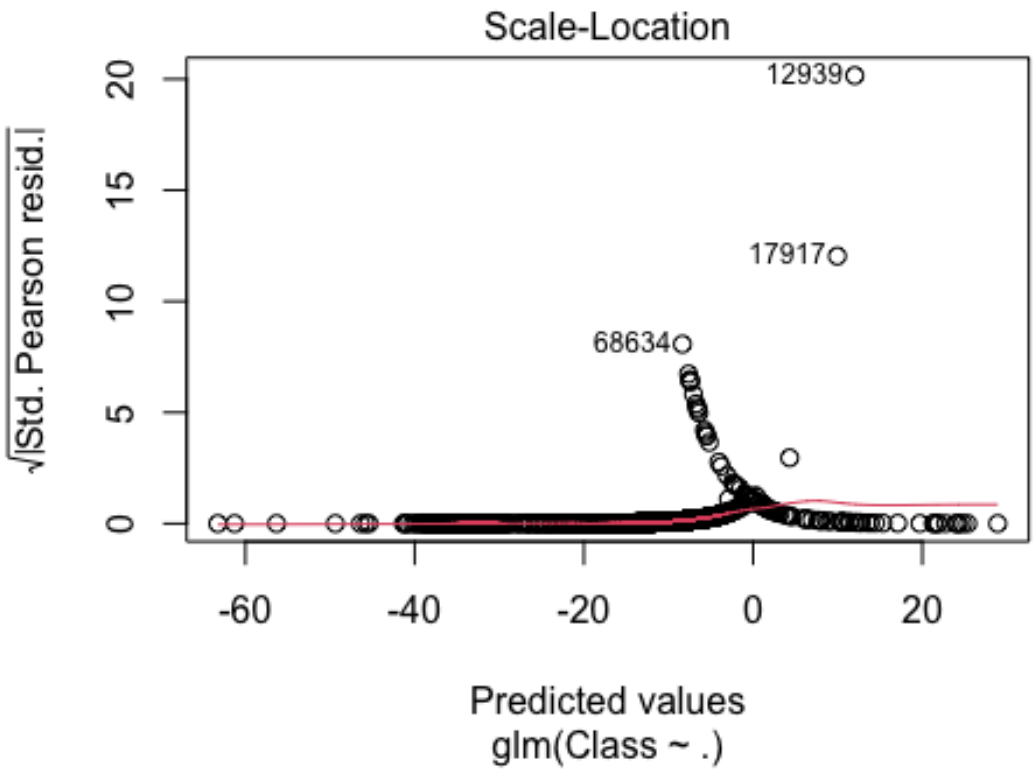
```
##
## Call:
## glm(formula = Class ~ ., family = binomial(), data = test_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.9019  -0.0254  -0.0156  -0.0078   4.0877
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.52800   10.30537  -1.216   0.2241
## V1           -0.17299    1.27381  -0.136   0.8920
## V2            1.44512    4.23062   0.342   0.7327
## V3            0.17897    0.24058   0.744   0.4569
## V4            3.13593    7.17768   0.437   0.6622
## V5            1.49014    3.80369   0.392   0.6952
## V6           -0.12428    0.22202  -0.560   0.5756
## V7            1.40903    4.22644   0.333   0.7388
## V8           -0.35254    0.17462  -2.019   0.0435 *
## V9            3.02176    8.67262   0.348   0.7275
## V10          -2.89571    6.62383  -0.437   0.6620
## V11          -0.09769    0.28270  -0.346   0.7297
## V12           1.97992    6.56699   0.301   0.7630
## V13          -0.71674    1.25649  -0.570   0.5684
## V14           0.19316    3.28868   0.059   0.9532
## V15           1.03868    2.89256   0.359   0.7195
## V16          -2.98194    7.11391  -0.419   0.6751
## V17          -1.81809    4.99764  -0.364   0.7160
## V18           2.74772    8.13188   0.338   0.7354
## V19          -1.63246    4.77228  -0.342   0.7323
## V20          -0.69925    1.15114  -0.607   0.5436
## V21          -0.45082    1.99182  -0.226   0.8209
## V22          -1.40395    5.18980  -0.271   0.7868
## V23           0.19026    0.61195   0.311   0.7559
## V24          -0.12889    0.44701  -0.288   0.7731
## V25          -0.57835    1.94988  -0.297   0.7668
## V26           2.65938    9.34957   0.284   0.7761
## V27          -0.45396    0.81502  -0.557   0.5775
```

```
## V28          -0.06639    0.35730  -0.186   0.8526
## Amount        0.22576    0.71892   0.314   0.7535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1443.40  on 56960  degrees of freedom
## Residual deviance:  378.59  on 56931  degrees of freedom
## AIC: 438.59
##
## Number of Fisher Scoring iterations: 17
```

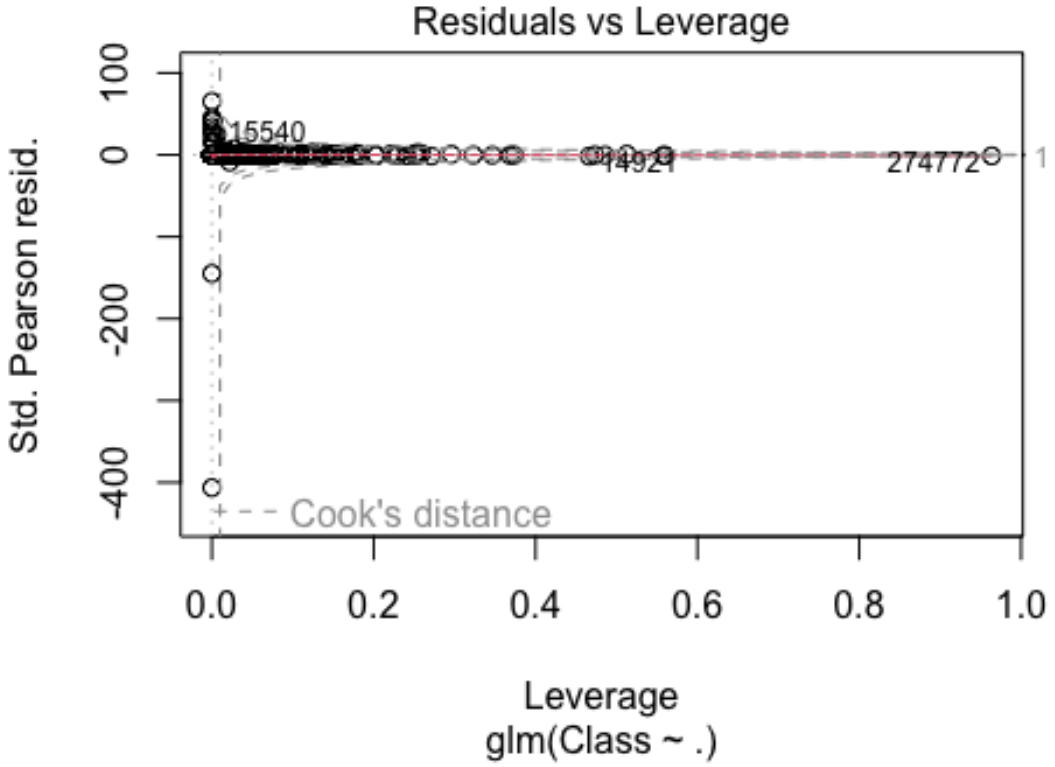
PLOTTING we also plot the regression model below and we see different plots first plot is residual vs fitted values which shows the predicted value is normal second plot shows the normality of the predicted values which look normal and the straight constant line is observed, 3rd plot is the scale location of the plot and we have seen all the predicted value of the data is near to the scale value which means there is no fraudelent activity is visible in any transaction there is no missing value so the 4th plot is not applicable and the 5th plot is the plott which observed the residuals value vs leverage as the value are near to each other so we have not see leverages

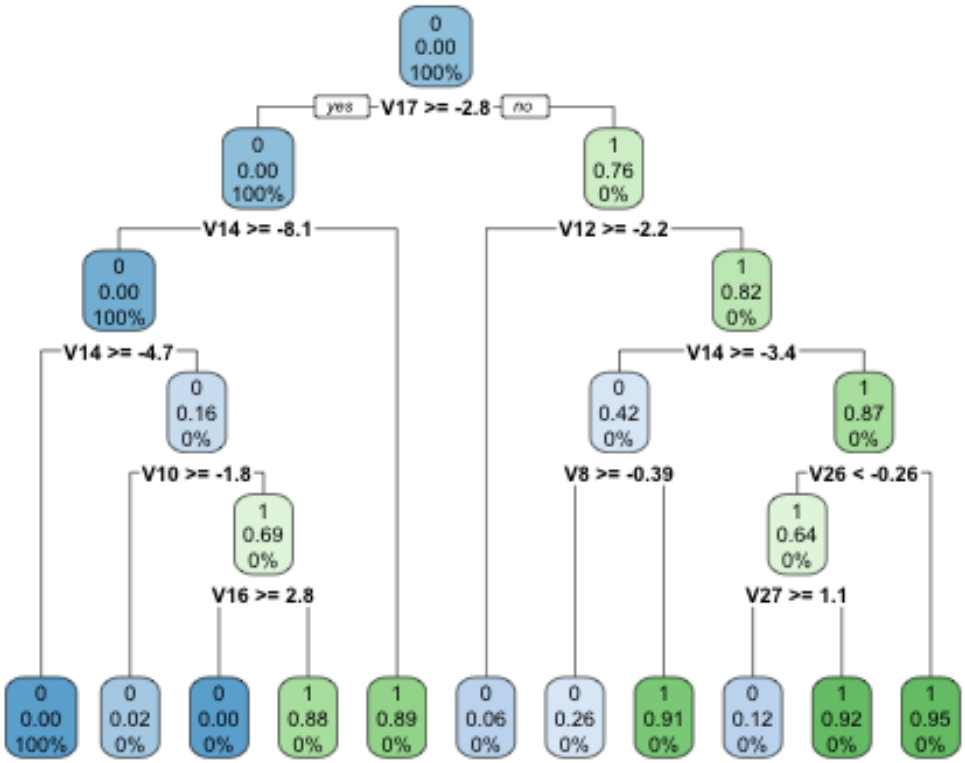

Residuals vs Fitted
Residuals
100
0
-200
-400
68634
17917
12939
-60
-40
-20
0
20
Predicted values
glm(Class ~ .)


Normal Q-Q
Std. Pearson resid.
100
0
-200
-400
68634
17917
12939
-4
-2
0
2
4
Theoretical Quantiles
glm(Class ~ .)




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```


Residuals vs Leverage
Std. Pearson resid.
100
0
-200
-400
15540
14921
274772
1
Cook's distance
0.0
0.2
0.4
0.6
0.8
1.0
Leverage
glm(Class ~ .)



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.