STA6714-19Spring 0001- Term Project Step 2--Data Analysis

Mithun Mohanraj

Task : Binary classification on predicting the rain

Data pre-processing:

The rain dataset has a lot of missing data for few predictors .so I have to drop all the rows that has missing data .

R code for missing data:

```
new.data <- na.omit(rain)
```

Also , date variable is trivial and had to be dropped

```
new.data$Date= NULL
```

The rain dataset has 6 categorical variables . They have to be converted into numerical factor variables before fitting the model.

R code for factors:

```
new.data$Location=as.numeric(factor(new.data$Location))
new.data$WindGustDir=as.numeric(factor(new.data$WindGustDir))
new.data$WindDir9am=as.numeric(factor(new.data$WindDir9am))
new.data$WindDir3pm=as.numeric(factor(new.data$WindDir3pm))
new.data$RainToday=as.numeric(factor(new.data$RainToday))
new.data$RainTomorrow=as.numeric(factor(new.data$RainTomorrow))
```

Logistic regression :

The logistic regression model is fitted on pre-processed data .

R code :

```
logit.reg <- glm(new.data$RainTomorrow~ ., data = new.data[,-23], family =
"binomial",control = list(maxit = 50, epsilon=1))
```

options(scipen=999)

summary(logit.reg)

Results are:

```
Call:
glm(formula = new.data$RainTomorrow ~ ., family = "binomial",
    data = new.data[, -23], control = list(maxit = 50, epsilon = 1))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1999  -0.6110  -0.4635  -0.3310   2.5036

Coefficients:
                Estimate Std. Error z value             Pr(>|z|)
(Intercept)   32.9911949  1.9770381  16.687 < 0.0000000000000002 ***
Location      -0.0009411  0.0014116  -0.667             0.504950
MinTemp       -0.0350366  0.0050286  -6.967      0.00000000000323 ***
MaxTemp        0.0418879  0.0095859   4.370      0.00001243868872 ***
Rainfall      -0.0052146  0.0017245  -3.024             0.002497 **
Evaporation    0.0075358  0.0039671   1.900             0.057492 .
Sunshine      -0.1018064  0.0047453 -21.454 < 0.0000000000000002 ***
WindGustDir    0.0093092  0.0027857   3.342             0.000832 ***
WindGustSpeed  0.0304567  0.0012554  24.260 < 0.0000000000000002 ***
WindDir9am    -0.0083984  0.0024729  -3.396             0.000683 ***
WindDir3pm     0.0046143  0.0027376   1.686             0.091892 .
WindSpeed9am  -0.0023658  0.0016050  -1.474             0.140475
WindSpeed3pm  -0.0181122  0.0016944 -10.690 < 0.0000000000000002 ***
Humidity9am   -0.0033400  0.0011128  -3.001             0.002688 **
Humidity3pm    0.0320137  0.0012756  25.098 < 0.0000000000000002 ***
Pressure9am    0.0627524  0.0068404   9.174 < 0.0000000000000002 ***
Pressure3pm   -0.0991025  0.0067687 -14.641 < 0.0000000000000002 ***
Cloud9am      -0.0202418  0.0052046  -3.889             0.000101 ***
Cloud3pm       0.0263413  0.0055461   4.750      0.00000203898403 ***
Temp9am       -0.0043457  0.0073643  -0.590             0.555123
Temp3pm       -0.0132852  0.0106769  -1.244             0.213389
RainToday      0.4390677  0.0310329  14.148 < 0.0000000000000002 ***
RISK_MM        0.0677015  0.0012621  53.643 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59493  on 56419  degrees of freedom
Residual deviance: 35641  on 56397  degrees of freedom
AIC: 35687
```

As we see above , there are few number of predictors which are statistically insignificant having high p-values > 0.05 for predicting the response variable.

Hence we run stepwise logistic regression on the pre-processed data.

Confusion matrix :

```
    pred
        0     1
No  43171   822
Yes  6132  6295
```

ACCURACY : 87%

Step-wise logistic regression :

Results are :

```
glm(formula = new.data$RainTomorrow ~ MinTemp + MaxTemp + Rainfall +
    Sunshine + WindGustDir + WindGustSpeed + WindDir9am + WindDir3pm +
    WindSpeed9am + WindSpeed3pm + Humidity9am + Humidity3pm +
    Pressure9am + Pressure3pm + Cloud9am + Cloud3pm + RainToday +
    RISK_MM, family = "binomial", data = new.data[, -23], control = list(maxit = 50,
    epsilon = 1))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1812  -0.6116  -0.4635  -0.3301   2.5159

Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
(Intercept)  33.0571032  1.9669704  16.806 < 0.0000000000000002 ***
MinTemp      -0.0376434  0.0036525 -10.306 < 0.0000000000000002 ***
MaxTemp       0.0300431  0.0039321   7.640   0.0000000000000217 ***
Rainfall     -0.0051523  0.0017189  -2.997             0.002722 **
Sunshine     -0.1016516  0.0047279 -21.500 < 0.0000000000000002 ***
WindGustDir   0.0094464  0.0027767   3.402             0.000669 ***
WindGustSpeed 0.0309821  0.0012305  25.178 < 0.0000000000000002 ***
WindDir9am   -0.0081792  0.0024578  -3.328             0.000875 ***
WindDir3pm    0.0046593  0.0027370   1.702             0.088692 .
WindSpeed9am -0.0021778  0.0015791  -1.379             0.167833
WindSpeed3pm -0.0185729  0.0016519 -11.243 < 0.0000000000000002 ***
Humidity9am  -0.0037030  0.0008311  -4.455   0.0000083700372431 ***
Humidity3pm   0.0322340  0.0009495  33.947 < 0.0000000000000002 ***
Pressure9am   0.0595823  0.0065645   9.076 < 0.0000000000000002 ***
Pressure3pm  -0.0960221  0.0064983 -14.777 < 0.0000000000000002 ***
Cloud9am     -0.0195747  0.0051831  -3.777             0.000159 ***
Cloud3pm      0.0268633  0.0055246   4.863   0.0000011591123352 ***
RainToday     0.4312313  0.0308423  13.982 < 0.0000000000000002 ***
RISK_MM       0.0677434  0.0012618  53.689 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59493  on 56419  degrees of freedom
Residual deviance: 35637  on 56401  degrees of freedom
AIC: 35675

Number of Fisher Scoring iterations: 1
```

As we see above , the number of predictors(18 out of total 23) is less after running step-wise regression .

Also the null deviance is 59493 on 56419 df is reduced to 35637 on 56401 when using predictors which indicates the model is a good fit.

Using the important predictors via stepwise model results in almost same residual deviance as that of model using all the predictors.

The predictors used are :

```
MinTemp + MaxTemp + Rainfall + Sunshine + WindGustDir + WindGustSpeed +
WindDir9am + WindDir3pm + WindSpeed9am + WindSpeed3pm + Humidity9am +
Humidity3pm + Pressure9am + Pressure3pm + Cloud9am + Cloud3pm + RainToday
+ RISK_MM

These predictors are statistically significant as their p-value< 0.05.
```

Confusion matrix:

```
     pred
        0     1
No  43178   815
Yes  6139  6288
```

Accuracy : 87%

R code :

```
stepw1=step(logit.reg,trace = 0)

summary(stepw1)
```

To ensure that these variables are statistically important , we can run chi-square test between the response variable and predictors :

Chi-square test results :

For mintemp:

```
chisq.test(new.data$RainTomorrow,new.data$MinTemp,correct = FALSE)
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$MinTemp
X-squared = 1277.8, df = 347, p-value < 0.00000000000000022
```

For maxtemp:

```
chisq.test(new.data$RainTomorrow,new.data$MaxTemp,correct = FALSE)
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$MaxTemp
X-squared = 1918, df = 394, p-value < 0.00000000000000022
```

For rainfall:

```
chisq.test(new.data$RainTomorrow,new.data$Rainfall,correct = FALSE)
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$Rainfall
X-squared = 7351.7, df = 409, p-value < 0.00000000000000022
```

For sunshine :

```
chisq.test(new.data$RainTomorrow,new.data$Sunshine,correct = FALSE)
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$Sunshine
X-squared = 12103, df = 144, p-value < 0.00000000000000022
```

FOR WINDGUSTSPEED:

```
chisq.test(new.data$RainTomorrow,new.data$WindGustSpeed,correct = FALSE)
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$WindGustSpeed
X-squared = 3405.1, df = 60, p-value < 0.00000000000000022
```

Fr winddir9am:

```
chisq.test(new.data$RainTomorrow,new.data$WindDir9am,correct = FALSE)
```

```
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$WindDir9am
X-squared = 987.91, df = 15, p-value < 0.00000000000000022
```

For windspeed9am:

```
chisq.test(new.data$RainTomorrow,new.data$WindSpeed9am,correct = FALSE)
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$WindSpeed9am
X-squared = 482.54, df = 35, p-value < 0.00000000000000022
```

For humidity9am:

```
chisq.test(new.data$RainTomorrow,new.data$Humidity9am,correct = FALSE)
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$Humidity9am
X-squared = 5124.5, df = 100, p-value < 0.00000000000000022
```

For raintoday:

```
chisq.test(new.data$RainTomorrow,new.data$RainToday,correct = FALSE)
```

```
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$RainToday
X-squared = 5390.5, df = 1, p-value < 0.00000000000000022
```

For riskmm:

```
chisq.test(new.data$RainTomorrow,new.data$RISK_MM,correct = FALSE)
```

```
Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  new.data$RainTomorrow and new.data$RISK_MM
X-squared = 56420, df = 497, p-value < 0.00000000000000022
```

Since all the predictors obtained through stepwise model have chi-square p-value less than 0.05 obtained by chi-square test between response and predictors , response variable is dependent on those variables .

Statistically significant variables:

MinTemp + MaxTemp + Rainfall + Sunshine + WindGustDir + WindGustSpeed + WindDir9am + WindDir3pm + WindSpeed9am + WindSpeed3pm + Humidity9am + Humidity3pm + Pressure9am + Pressure3pm + Cloud9am + Cloud3pm + RainToday + RISK_MM

PCA:

PCA cannot be conducted because most of the variables are categorical having more than 2 classes.