logistic_regression.R

Magilan

Mon Oct 08 16:36:47 2018

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(boot)
library(forecast)
library(tseries)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:boot':
##
##     melanoma
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(psych)
```

```
##
## Attaching package: 'psych'
## The following object is masked from 'package:boot':
## 
##     logit
```

```
## The following objects are masked from 'package:ggplot2':
## 
##     %+%, alpha
```

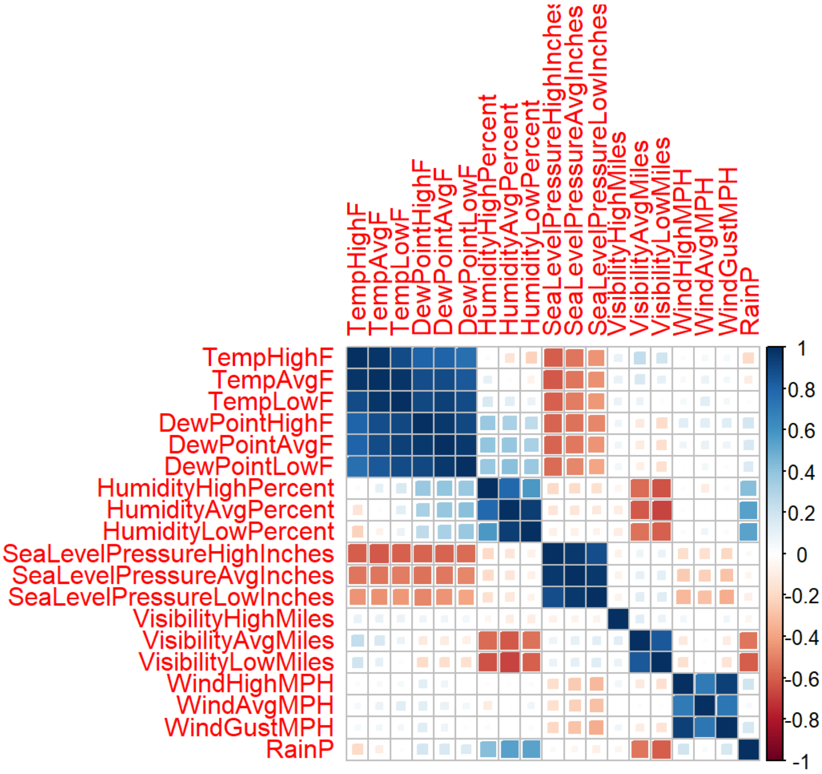
```
# Data Input

data <- read.csv("C:/Users/Magilan/Desktop/ML_project/austin_weather.csv",header = TRUE)
data1=na.omit(data,invert=FALSE)
attach(data1)
summary(data1)
```

```
##          Date        TempHighF         TempAvgF        TempLowF    
##  01-01-2014:   1   Min.   : 32.00   Min.   :29.00   Min.   :19.00  
##  01-01-2015:   1   1st Qu.: 72.00   1st Qu.:62.00   1st Qu.:49.00  
##  01-02-2014:   1   Median : 83.00   Median :73.00   Median :62.00  
##  01-02-2015:   1   Mean   : 80.79   Mean   :70.56   Mean   :59.82  
##  01-02-2016:   1   3rd Qu.: 92.00   3rd Qu.:83.00   3rd Qu.:73.00  
##  01-02-2017:   1   Max.   :107.00   Max.   :93.00   Max.   :81.00  
##  (Other)   :1299                                                   
##  DewPointHighF    DewPointAvgF    DewPointLowF   HumidityHighPercent
##  Min.   :13.00   Min.   : 8.00   Min.   : 2.00   Min.   : 37.00     
##  1st Qu.:53.00   1st Qu.:46.00   1st Qu.:38.00   1st Qu.: 85.00     
##  Median :66.00   Median :61.00   Median :56.00   Median : 90.00     
##  Mean   :61.52   Mean   :56.64   Mean   :50.94   Mean   : 87.83     
##  3rd Qu.:73.00   3rd Qu.:69.00   3rd Qu.:65.00   3rd Qu.: 94.00     
##  Max.   :80.00   Max.   :76.00   Max.   :75.00   Max.   :100.00     
##                                                                     
##  HumidityAvgPercent HumidityLowPercent SeaLevelPressureHighInches
##  Min.   :27.00      Min.   :10.00      Min.   :29.63             
##  1st Qu.:59.00      1st Qu.:33.00      1st Qu.:29.99             
##  Median :67.00      Median :44.00      Median :30.08             
##  Mean   :66.66      Mean   :44.98      Mean   :30.11             
##  3rd Qu.:74.00      3rd Qu.:55.00      3rd Qu.:30.21             
##  Max.   :97.00      Max.   :93.00      Max.   :30.83             
##                                                                  
##  SeaLevelPressureAvgInches SeaLevelPressureLowInches VisibilityHighMiles
##  Min.   :29.55             Min.   :29.41             Min.   : 5.000     
##  1st Qu.:29.91             1st Qu.:29.82             1st Qu.:10.000     
##  Median :30.00             Median :29.91             Median :10.000     
##  Mean   :30.02             Mean   :29.93             Mean   : 9.992     
##  3rd Qu.:30.10             3rd Qu.:30.02             3rd Qu.:10.000     
##  Max.   :30.74             Max.   :30.61             Max.   :10.000     
##                                                                         
##  VisibilityAvgMiles VisibilityLowMiles  WindHighMPH      WindAvgMPH    
##  Min.   : 2.000     Min.   : 0.000     Min.   : 6.00   Min.   : 1.000  
##  1st Qu.: 9.000     1st Qu.: 3.000     1st Qu.:10.00   1st Qu.: 3.000  
##  Median :10.000     Median : 9.000     Median :13.00   Median : 5.000  
##  Mean   : 9.162     Mean   : 6.843     Mean   :13.25   Mean   : 5.009  
##  3rd Qu.:10.000     3rd Qu.:10.000     3rd Qu.:15.00   3rd Qu.: 6.000  
##  Max.   :10.000     Max.   :10.000     Max.   :29.00   Max.   :12.000  
##                                                                        
##   WindGustMPH    PrecipitationSumInches  Rain          RainP       
##  Min.   : 9.00   Min.   :0.0000         no :859   Min.   :0.0000  
##  1st Qu.:17.00   1st Qu.:0.0000         yes:446   1st Qu.:0.0000  
##  Median :21.00   Median :0.0000                   Median :0.0000  
##  Mean   :21.38   Mean   :0.1248                   Mean   :0.3418  
##  3rd Qu.:25.00   3rd Qu.:0.0800                   3rd Qu.:1.0000  
##  Max.   :57.00   Max.   :5.2000                   Max.   :1.0000  
## 
```

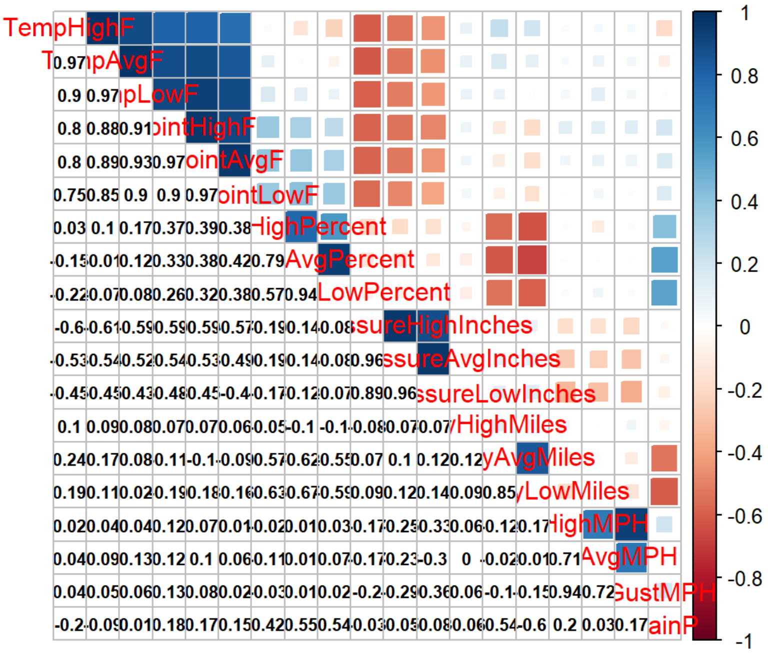
```
summary(Rain)
```

```
##  no yes 
## 859 446
```

```
mat=cor(data1[,-c(1,20,21)],method = "spearman")


corrplot(mat,method = "square")
```



```
corrplot.mixed(mat, lower.col = "black",upper = "square", number.cex = .7)
```



```r
# Data Partitioning

index <- createDataPartition(Rain, p = 0.7, list = FALSE)
# Training set
train.df <- data1[index,]
# Testing dataset
test.df <- data1[-index,]

summary(train.df)
```

```
##          Date       TempHighF         TempAvgF        TempLowF
##  01-01-2014:  1   Min.   : 32.00   Min.   :29.00   Min.   :19.00
##  01-01-2015:  1   1st Qu.: 71.00   1st Qu.:61.00   1st Qu.:49.00
##  01-02-2014:  1   Median : 83.00   Median :73.00   Median :62.00
##  01-02-2015:  1   Mean   : 80.62   Mean   :70.45   Mean   :59.78
##  01-03-2014:  1   3rd Qu.: 92.00   3rd Qu.:83.00   3rd Qu.:73.00
##  01-03-2015:  1   Max.   :107.00   Max.   :93.00   Max.   :80.00
##  (Other)   :909
##  DewPointHighF    DewPointAvgF    DewPointLowF   HumidityHighPercent
##  Min.   :13.00   Min.   :11.00   Min.   : 4.00   Min.   : 37.00
##  1st Qu.:52.00   1st Qu.:46.00   1st Qu.:38.00   1st Qu.: 84.50
##  Median :66.00   Median :61.00   Median :56.00   Median : 90.00
##  Mean   :61.35   Mean   :56.52   Mean   :50.87   Mean   : 87.82
##  3rd Qu.:73.00   3rd Qu.:69.00   3rd Qu.:65.00   3rd Qu.: 94.00
##  Max.   :80.00   Max.   :76.00   Max.   :75.00   Max.   :100.00
##
##  HumidityAvgPercent HumidityLowPercent SeaLevelPressureHighInches
##  Min.   :27.00      Min.   :10         Min.   :29.63
##  1st Qu.:59.00      1st Qu.:32         1st Qu.:30.00
##  Median :67.00      Median :44         Median :30.08
##  Mean   :66.68      Mean   :45         Mean   :30.12
##  3rd Qu.:75.00      3rd Qu.:55         3rd Qu.:30.21
##  Max.   :97.00      Max.   :93         Max.   :30.83
##
##  SeaLevelPressureAvgInches SeaLevelPressureLowInches VisibilityHighMiles
##  Min.   :29.55             Min.   :29.42             Min.   : 8.000
##  1st Qu.:29.92             1st Qu.:29.83             1st Qu.:10.000
##  Median :30.00             Median :29.92             Median :10.000
##  Mean   :30.03             Mean   :29.94             Mean   : 9.993
##  3rd Qu.:30.11             3rd Qu.:30.02             3rd Qu.:10.000
##  Max.   :30.74             Max.   :30.61             Max.   :10.000
##
##  VisibilityAvgMiles VisibilityLowMiles  WindHighMPH      WindAvgMPH
##  Min.   : 2.000     Min.   : 0.000     Min.   : 7.00   Min.   : 1.000
##  1st Qu.: 9.000     1st Qu.: 3.000     1st Qu.:10.00   1st Qu.: 3.000
##  Median :10.000     Median : 9.000     Median :13.00   Median : 5.000
##  Mean   : 9.158     Mean   : 6.902     Mean   :13.23   Mean   : 5.019
##  3rd Qu.:10.000     3rd Qu.:10.000     3rd Qu.:15.00   3rd Qu.: 6.000
##  Max.   :10.000     Max.   :10.000     Max.   :29.00   Max.   :11.000
##
##   WindGustMPH    PrecipitationSumInches  Rain         RainP
##  Min.   : 9.00   Min.   :0.0000         no :602   Min.   :0.0000
##  1st Qu.:17.00   1st Qu.:0.0000         yes:313   1st Qu.:0.0000
##  Median :21.00   Median :0.0000                   Median :0.0000
##  Mean   :21.38   Mean   :0.1164                   Mean   :0.3421
##  3rd Qu.:25.00   3rd Qu.:0.0800                   3rd Qu.:1.0000
##  Max.   :57.00   Max.   :4.9300                   Max.   :1.0000
##
```

```r
summary(test.df)
```

```
##          Date       TempHighF        TempAvgF        TempLowF    
##  01-02-2016:  1   Min.   : 36.0   Min.   :29.00   Min.   :22.00  
##  01-02-2017:  1   1st Qu.: 73.0   1st Qu.:62.00   1st Qu.:51.00  
##  01-05-2016:  1   Median : 83.0   Median :73.00   Median :62.00  
##  01-08-2016:  1   Mean   : 81.2   Mean   :70.81   Mean   :59.92  
##  01-10-2015:  1   3rd Qu.: 92.0   3rd Qu.:82.00   3rd Qu.:72.00  
##  01-10-2016:  1   Max.   :104.0   Max.   :92.00   Max.   :81.00  
##  (Other)   :384                                                  
##  DewPointHighF    DewPointAvgF    DewPointLowF   HumidityHighPercent
##  Min.   :15.00   Min.   : 8.00   Min.   : 2.00   Min.   : 44.00     
##  1st Qu.:54.25   1st Qu.:47.00   1st Qu.:38.00   1st Qu.: 85.00     
##  Median :66.00   Median :61.00   Median :55.00   Median : 91.00     
##  Mean   :61.90   Mean   :56.91   Mean   :51.13   Mean   : 87.86     
##  3rd Qu.:73.00   3rd Qu.:69.75   3rd Qu.:65.00   3rd Qu.: 94.00     
##  Max.   :78.00   Max.   :74.00   Max.   :73.00   Max.   :100.00     
##                                                                     
##  HumidityAvgPercent HumidityLowPercent SeaLevelPressureHighInches
##  Min.   :27.00      Min.   :10.00      Min.   :29.65             
##  1st Qu.:60.00      1st Qu.:33.00      1st Qu.:29.99             
##  Median :67.00      Median :44.00      Median :30.08             
##  Mean   :66.62      Mean   :44.94      Mean   :30.10             
##  3rd Qu.:74.00      3rd Qu.:54.00      3rd Qu.:30.19             
##  Max.   :97.00      Max.   :93.00      Max.   :30.80             
##                                                                  
##  SeaLevelPressureAvgInches SeaLevelPressureLowInches VisibilityHighMiles
##  Min.   :29.56             Min.   :29.41             Min.   : 5.000     
##  1st Qu.:29.91             1st Qu.:29.81             1st Qu.:10.000     
##  Median :30.00             Median :29.91             Median :10.000     
##  Mean   :30.01             Mean   :29.92             Mean   : 9.987     
##  3rd Qu.:30.10             3rd Qu.:30.01             3rd Qu.:10.000     
##  Max.   :30.68             Max.   :30.50             Max.   :10.000     
##                                                                         
##  VisibilityAvgMiles VisibilityLowMiles  WindHighMPH      WindAvgMPH    
##  Min.   : 2.000     Min.   : 0.000     Min.   : 6.00   Min.   : 1.000  
##  1st Qu.: 9.000     1st Qu.: 2.000     1st Qu.:10.00   1st Qu.: 3.000  
##  Median :10.000     Median : 9.000     Median :13.00   Median : 5.000  
##  Mean   : 9.172     Mean   : 6.705     Mean   :13.28   Mean   : 4.987  
##  3rd Qu.:10.000     3rd Qu.:10.000     3rd Qu.:15.00   3rd Qu.: 6.000  
##  Max.   :10.000     Max.   :10.000     Max.   :25.00   Max.   :12.000  
##                                                                        
##   WindGustMPH   PrecipitationSumInches  Rain         RainP      
##  Min.   : 9.0   Min.   :0.0000         no :257   Min.   :0.000  
##  1st Qu.:17.0   1st Qu.:0.0000         yes:133   1st Qu.:0.000  
##  Median :21.0   Median :0.0000                   Median :0.000  
##  Mean   :21.4   Mean   :0.1445                   Mean   :0.341  
##  3rd Qu.:25.0   3rd Qu.:0.0600                   3rd Qu.:1.000  
##  Max.   :43.0   Max.   :5.2000                   Max.   :1.000  
## 
```

```
# Logistic regression

colnames(data1)
```

```
##  [1] "Date"                       "TempHighF"                 
##  [3] "TempAvgF"                   "TempLowF"                  
##  [5] "DewPointHighF"              "DewPointAvgF"              
##  [7] "DewPointLowF"               "HumidityHighPercent"       
##  [9] "HumidityAvgPercent"         "HumidityLowPercent"        
## [11] "SeaLevelPressureHighInches" "SeaLevelPressureAvgInches" 
## [13] "SeaLevelPressureLowInches"  "VisibilityHighMiles"       
## [15] "VisibilityAvgMiles"         "VisibilityLowMiles"        
## [17] "WindHighMPH"                "WindAvgMPH"                
## [19] "WindGustMPH"                "PrecipitationSumInches"    
## [21] "Rain"                       "RainP"
```

```
model <- glm(Rain ~ TempHighF+TempAvgF+TempLowF+DewPointHighF+DewPointAvgF+DewPointLowF+HumidityHighPercent+
HumidityAvgPercent+HumidityLowPercent+SeaLevelPressureHighInches+SeaLevelPressureAvgInches+VisibilityLowMile
s+VisibilityHighMiles+VisibilityAvgMiles+WindGustMPH+WindHighMPH+WindAvgMPH, data = train.df, family = binom
ial)

summary(model)
```

```
##
## Call:
## glm(formula = Rain ~ TempHighF + TempAvgF + TempLowF + DewPointHighF +
##     DewPointAvgF + DewPointLowF + HumidityHighPercent + HumidityAvgPercent +
##     HumidityLowPercent + SeaLevelPressureHighInches + SeaLevelPressureAvgInches +
##     VisibilityLowMiles + VisibilityHighMiles + VisibilityAvgMiles +
##     WindGustMPH + WindHighMPH + WindAvgMPH, family = binomial,
##     data = train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7676  -0.4454  -0.1951   0.3632   2.7014
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -40.99834   29.17451  -1.405   0.1599
## TempHighF                   -0.04739    0.20958  -0.226   0.8211
## TempAvgF                    -0.35795    0.41079  -0.871   0.3836
## TempLowF                     0.33203    0.20893   1.589   0.1120
## DewPointHighF                0.08541    0.03887   2.197   0.0280 *
## DewPointAvgF                 0.08855    0.06375   1.389   0.1648
## DewPointLowF                -0.05499    0.03454  -1.592   0.1114
## HumidityHighPercent         -0.09692    0.07935  -1.221   0.2219
## HumidityAvgPercent           0.13546    0.15352   0.882   0.3776
## HumidityLowPercent          -0.04748    0.07819  -0.607   0.5437
## SeaLevelPressureHighInches   2.36409    3.07149   0.770   0.4415
## SeaLevelPressureAvgInches   -1.10078    3.20509  -0.343   0.7313
## VisibilityLowMiles          -0.36397    0.05300  -6.868 6.52e-12 ***
## VisibilityHighMiles          0.22131    0.75491   0.293   0.7694
## VisibilityAvgMiles           0.29502    0.12463   2.367   0.0179 *
## WindGustMPH                  0.06403    0.05829   1.099   0.2719
## WindHighMPH                  0.25335    0.10057   2.519   0.0118 *
## WindAvgMPH                  -0.43272    0.08616  -5.023 5.10e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1175.60  on 914  degrees of freedom
## Residual deviance:  585.35  on 897  degrees of freedom
## AIC: 621.35
##
## Number of Fisher Scoring iterations: 6
```

```
predicted_values <- predict(model, test.df[,-c(1,20,21,22)], type = "response")
head(predicted_values)
```

```
##           1           3           6           9          15          16
## 0.868356607 0.004696376 0.296961362 0.676644194 0.090897401 0.769032689
```

```
# Validation

table(Rain)
```

```
## Rain
##  no yes
## 859 446
```

```
nrows_prediction<-nrow(test.df)
prediction <- data.frame(c(1:nrows_prediction))
colnames(prediction) <- c("Rain")
str(prediction)
```

```
## 'data.frame':    390 obs. of  1 variable:
##  $ Rain: int  1 2 3 4 5 6 7 8 9 10 ...
```

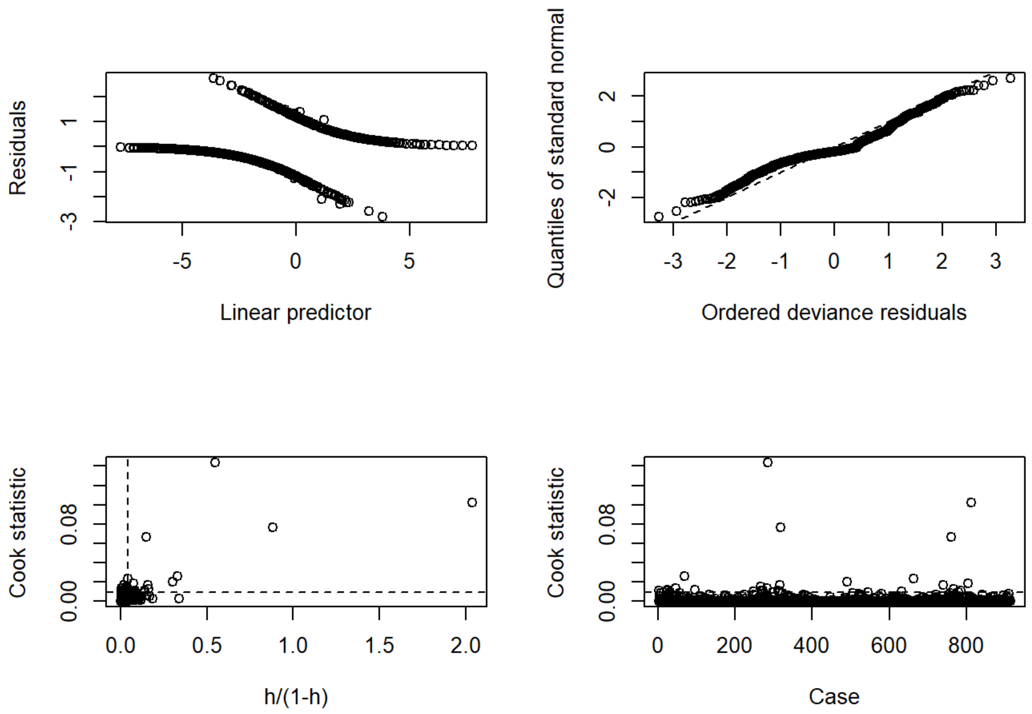
```
prediction$Rain <- as.character(prediction$Rain)
prediction$Rain <- "yes"
prediction$Rain[ predicted_values < 0.5] <- "no"
prediction$Rain <- as.factor(prediction$Rain)

#Confusion Matrix

table(prediction$Rain, test.df$Rain)
```

```
##      
##        no yes
##   no  230  31
##   yes  27 102
```

```
confusionMatrix(prediction$Rain,test.df$Rain)
```

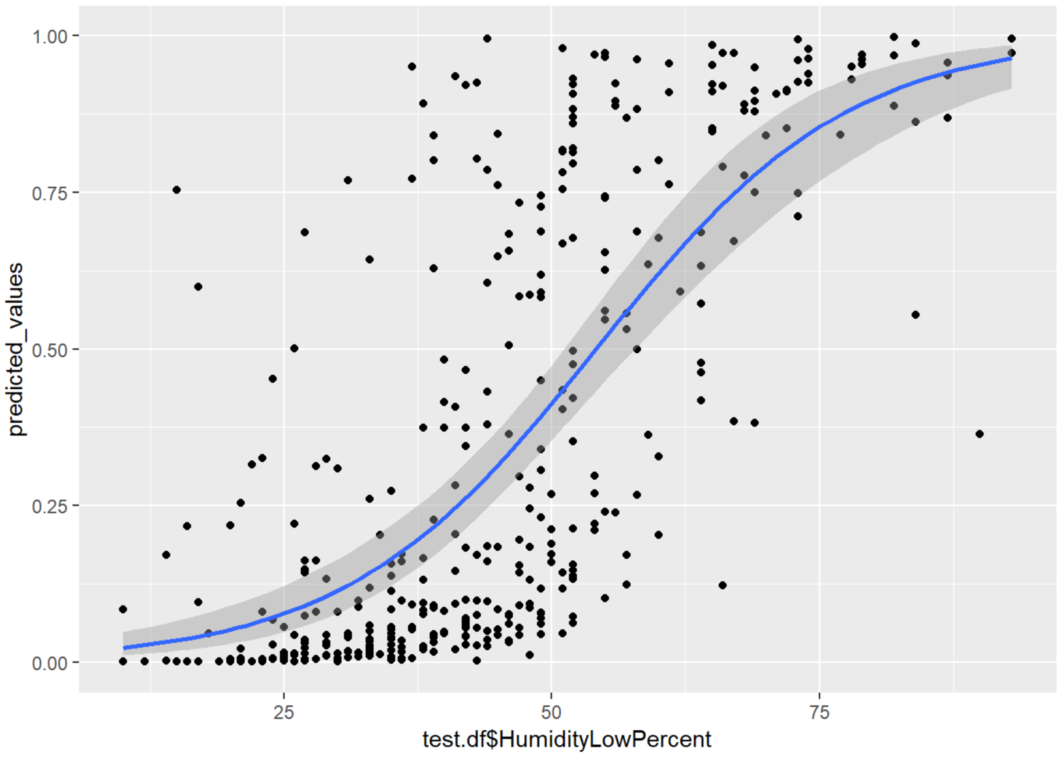
```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction  no yes
##        no  230  31
##        yes  27 102
##                                          
##                Accuracy : 0.8513         
##                  95% CI : (0.812, 0.8851)
##     No Information Rate : 0.659          
##     P-Value [Acc > NIR] : <2e-16         
##                                          
##                   Kappa : 0.6667         
##  Mcnemar's Test P-Value : 0.6936         
##                                          
##             Sensitivity : 0.8949         
##             Specificity : 0.7669         
##          Pos Pred Value : 0.8812         
##          Neg Pred Value : 0.7907         
##              Prevalence : 0.6590         
##          Detection Rate : 0.5897         
##    Detection Prevalence : 0.6692         
##       Balanced Accuracy : 0.8309         
##                                          
##        'Positive' Class : no             
## 
```

```
glm.diag.plots(model)
```



```
ggplot(test.df, aes(x = test.df$HumidityLowPercent, y = predicted_values))+
  geom_point() + # add points
  geom_smooth(method = "glm", # plot a regression...
              method.args = list(family = "binomial"))
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial
## glm!
```


predicted_values
test.df$HumidityLowPercent