logistic Reggresion

Maliheh_Garoosiha

2024-04-08

```
yelp.data<-read.csv("Yelp_dataset.csv",header=TRUE)
head(yelp.data)
```

```
##      city class review_count           categories
## 1 Toronto     0           12              Italian
## 2 Toronto     0           39                  Pub
## 3 Toronto     1            3 Coffee or Sandwiches
## 4 Toronto     1           55       Middle Eastern
## 5 Markham     0           80                Asian
## 6 Toronto     0            5                Asian
```

```
str(yelp.data)
```

```
## 'data.frame':    9219 obs. of  4 variables:
##  $ city        : chr  "Toronto" "Toronto" "Toronto" "Toronto" ...
##  $ class       : int  0 0 1 1 0 0 0 0 0 1 ...
##  $ review_count: int  12 39 3 55 80 5 6 6 34 8 ...
##  $ categories  : chr  "Italian" "Pub" "Coffee or Sandwiches" "Middle Eastern" ...
```

```
set.seed(123)
train.index<-sample(1:nrow(yelp.data), .7*nrow(yelp.data))
train.set<-yelp.data[train.index,]
test.set<-yelp.data[-train.index,]
glm_model<-glm(class~.,family="binomial",data=train.set)
summary(glm_model)
```

```
##
## Call:
## glm(formula = class ~ ., family = "binomial", data = train.set)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -1.5351114  0.1151796 -13.328  < 2e-16 ***
## cityMissisauga                 0.4881207  0.1318810   3.701 0.000215 ***
## cityToronto                    0.5460575  0.1142741   4.778 1.77e-06 ***
## review_count                   0.0030342  0.0004157   7.300 2.88e-13 ***
## categoriesCoffee or Sandwiches 0.7970987  0.0913172   8.729  < 2e-16 ***
## categoriesFast Food           -0.2621972  0.0928911  -2.823 0.004763 **
## categoriesItalian              0.3832537  0.1635196   2.344 0.019089 *
## categoriesLatin                0.4941062  0.1322714   3.736 0.000187 ***
## categoriesMiddle Eastern       0.3573337  0.0929445   3.845 0.000121 ***
## categoriesNorth American       0.2659344  0.1001373   2.656 0.007914 **
## categoriesOther                0.5930035  0.1032110   5.746 9.16e-09 ***
## categoriesPub                  0.3972647  0.1145871   3.467 0.000526 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8302.0  on 6452  degrees of freedom
## Residual deviance: 8048.8  on 6441  degrees of freedom
## AIC: 8072.8
##
## Number of Fisher Scoring iterations: 4
```

```
predict(glm_model, newdata=data.frame(city="Toronto",review_count=200,categories="Coffee or Sandwiches"))
```

```
##         1
## 0.6022585
```

```
predicted <- predict(glm_model, test.set,type="response")
head(predicted)
```

```
##         3         6         8        10        12        14
## 0.4544138 0.2741072 0.3307235 0.1451653 0.2380755 0.3802679
```

```
peredicted_final<-ifelse(predicted>.5,1,0)
table_final<-table(actual=test.set$class,predicted=peredicted_final)
```

```
accuracy_percent<-(sum(diag(table_final)))/nrow(test.set)
print(accuracy_percent)
```

```
## [1] 0.664859
```