We use logistic regression when dependent variable is categorical.

**DataSet 1**

train1\_1<- read.csv("C:\\Users\\Shalini\\Documents\\Information Retrival\\Assignment\\IRProjectData\\fakeTrainData1.csv")

train1\_2 <- read.csv("C:\\Users\\Shalini\\Documents\\Information Retrival\\Assignment\\IRProjectData\\realTrainData1.csv")

train1\_1$last <- as.numeric(0)

train1\_2$last <- as.numeric(1)

full\_train1 <- rbind(train1\_1,train1\_2)

> binomFit1 <- glm(full\_train1$last ~ ., data = full\_train1[,-c(1,12)],family = "binomial")

> summary(binomFit1)

Call:

glm(formula = full\_train1$last ~ ., family = "binomial", data = full\_train1[,

-c(1, 12)])

Deviance Residuals:

Min 1Q Median 3Q Max

-1.4029 -1.1765 0.3855 1.1703 1.6346

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 1.829e-02 1.564e-02 1.169 0.242226

trust 1.244e-01 1.603e-02 7.763 8.28e-15 \*\*\*

fear 1.371e-02 2.037e-02 0.673 0.500845

negative -1.441e-03 1.765e-02 -0.082 0.934938

sadness 3.274e-02 2.251e-02 1.454 0.145889

anger 4.861e-03 2.337e-02 0.208 0.835213

surprise 7.832e-02 2.245e-02 3.489 0.000485 \*\*\*

positive -5.313e-02 1.499e-02 -3.543 0.000395 \*\*\*

disgust 7.348e-05 2.438e-02 0.003 0.997595

joy -9.676e-02 1.942e-02 -4.982 6.30e-07 \*\*\*

anticipation -4.055e-02 1.619e-02 -2.505 0.012236 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 69315 on 49999 degrees of freedom

Residual deviance: 69180 on 49989 degrees of freedom

AIC: 69202

Number of Fisher Scoring iterations: 3

**Prediction using Test Data**

binomPredict <- predict(binomFit1,test1[,-c(1,12)],"response")

binomPredict[binomPredict > 0.59] <- 'fake'

binomPredict[binomPredict <= 0.59] <- 'real'

library(caret)

confusionMatrix(binomPredict,test1$realFAKEcat)

Confusion Matrix and Statistics

Reference

Prediction fake real

fake 1 23

real 536 14440

Accuracy : 0.9627

95% CI : (0.9596, 0.9657)

No Information Rate : 0.9642

P-Value [Acc > NIR] : 0.8387

Kappa : 5e-04

Mcnemar's Test P-Value : <2e-16

Sensitivity : 1.862e-03

Specificity : 9.984e-01

Pos Pred Value : 4.167e-02

Neg Pred Value : 9.642e-01

Prevalence : 3.580e-02

Detection Rate : 6.667e-05

Detection Prevalence : 1.600e-03

Balanced Accuracy : 5.001e-01

'Positive' Class : fake

**Data Set 2**

train2\_1 <- read.csv("C:\\Users\\Shalini\\Documents\\Information Retrival\\Assignment\\IRProjectData\\fakeTrainData2.csv")

train2\_2 <- read.csv("C:\\Users\\Shalini\\Documents\\Information Retrival\\Assignment\\IRProjectData\\realTrainData2.csv")

train2\_1$last <- as.numeric(0)

train2\_2$last <- as.numeric(1)

full\_train2 <- rbind(train2\_1,train2\_2)

binomFit2 <- glm(full\_train2$last ~ ., data = full\_train2[,-c(1,12)],family = "binomial")

summary(binomFit2)

Call:

glm(formula = full\_train2$last ~ ., family = "binomial", data = full\_train2[,

-c(1, 12)])

Deviance Residuals:

Min 1Q Median 3Q Max

-1.46252 -1.17636 0.09481 1.17590 1.45426

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.038419 0.034901 1.101 0.27098

trust 0.072272 0.035536 2.034 0.04197 \*

fear 0.023036 0.045926 0.502 0.61596

negative -0.040891 0.039879 -1.025 0.30519

sadness 0.119077 0.050783 2.345 0.01904 \*

anger -0.010581 0.052506 -0.202 0.84029

surprise 0.050141 0.049413 1.015 0.31023

positive -0.056097 0.032832 -1.709 0.08752 .

disgust 0.042491 0.055312 0.768 0.44236

joy -0.132719 0.043737 -3.034 0.00241 \*\*

anticipation 0.006022 0.035715 0.169 0.86610

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 13863 on 9999 degrees of freedom

Residual deviance: 13825 on 9989 degrees of freedom

AIC: 13847

Number of Fisher Scoring iterations: 3

**Prediction using Test Data**

binomPredict2 <- predict(binomFit2,test1[,-c(1,12)],"response")

summary(binomPred)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.2198 0.4827 0.5009 0.5015 0.5189 0.7509

binomPredict2[binomPredict2 > 0.5015] <- 'fake'

binomPredict2[binomPredict2 <= 0.5015] <- 'real'

library(caret)

confusionMatrix(binomPredict2,test1$realFAKEcat)

Confusion Matrix and Statistics

Reference

Prediction fake real

fake 238 7105

real 299 7358

Accuracy : 0.5064

95% CI : (0.4984, 0.5144)

No Information Rate : 0.9642

P-Value [Acc > NIR] : 1

Kappa : -0.0068

Mcnemar's Test P-Value : <2e-16

Sensitivity : 0.44320

Specificity : 0.50875

Pos Pred Value : 0.03241

Neg Pred Value : 0.96095

Prevalence : 0.03580

Detection Rate : 0.01587

Detection Prevalence : 0.48953

Balanced Accuracy : 0.47597

'Positive' Class : fake