Logistic_Regression.R

```r
#Top Songs Analysis

#importing dataset top10s and copying it to test data
data = read.csv('C:\\Users\\ApurvaSarode\\Desktop\\Spotify_mva.csv')

View(data)

#Data Cleaning

y = data$pop
View(y)
max(y)

## [1] 99

mean(y)

## [1] 66.52073

max(y)

## [1] 99

rating <- cut(y, breaks = c(0,67,99),
              labels = c("Below Average", "Above Average"),
              right = FALSE, include.lowest = TRUE)
data['rating'] <- rating
View(data)
data_clean <- data[-c(433),]
View(data_clean)

#set.seed(422)
#split = sample.split(data_clean, SplitRatio=0.8)
#train = subset(data_clean, split == TRUE)
#test = subset(data_clean, split == FALSE)
#dim(train)
#View(train)
#View(test)

library(ggplot2)
fit_lg <- glm(rating~nrgy+dB+dur,data = data_clean, family = "binomial")
summary(fit_lg)

##
## Call:
## glm(formula = rating ~ nrgy + dB + dur, family = "binomial",
##     data = data_clean)
```

```
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.6612  -1.2840   0.9274   1.0368   1.3749  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)   
## (Intercept)  3.126867   0.956132   3.270  0.00107 **
## nrgy        -0.017119   0.007002  -2.445  0.01449 * 
## dB           0.115497   0.063324   1.824  0.06817 . 
## dur         -0.004122   0.002488  -1.656  0.09766 . 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 816.49  on 601  degrees of freedom
## Residual deviance: 806.25  on 598  degrees of freedom
## AIC: 814.25
## 
## Number of Fisher Scoring iterations: 4
```

```
## Now calculate the overall "Pseudo R-squared" and its p-value
ll.null <- fit_lg$null.deviance/-2
ll.proposed <- fit_lg$deviance/-2
## McFadden's Pseudo R^2 = [ LL(Null) - LL(Proposed) ] / LL(Null)
(ll.null - ll.proposed) / ll.null
```

```
## [1] 0.01254404
```

```
## The p-value for the R^2
1 - pchisq(2*(ll.proposed - ll.null), df=(length(fit_lg$coefficients)-1))
```

```
## [1] 0.01661637
```

```
lrm <- glm(rating~nrgy+dnce+dB+val+dur+spch,data = data_clean, family =
"binomial")
summary(lrm)
```

```
## 
## Call:
## glm(formula = rating ~ nrgy + dnce + dB + val + dur + spch, family =
"binomial", 
##     data = data_clean)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8376  -1.2511   0.8648   1.0366   1.4465  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.820856   1.087954   1.674 0.094199 .
## nrgy        -0.016671   0.007427  -2.245 0.024792 *
## dnce         0.026445   0.007693   3.438 0.000587 ***
## dB           0.132145   0.067972   1.944 0.051883 .
## val         -0.009256   0.004904  -1.887 0.059102 .
## dur         -0.003869   0.002580  -1.500 0.133676
## spch         0.011611   0.011764   0.987 0.323620
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 816.49  on 601  degrees of freedom
## Residual deviance: 793.67  on 595  degrees of freedom
## AIC: 807.67
##
## Number of Fisher Scoring iterations: 4
```

```
## Now calculate the overall "Pseudo R-squared" and its p-value
ll.null <- lrm$null.deviance/-2
ll.proposed <- lrm$deviance/-2
## McFadden's Pseudo R^2 = [ LL(Null) - LL(Proposed) ] / LL(Null)
(ll.null - ll.proposed) / ll.null
```

```
## [1] 0.02795105
```

```
## The p-value for the R^2
1 - pchisq(2*(ll.proposed - ll.null), df=(length(lrm$coefficients)-1))
```

```
## [1] 0.0008584091
```

```
#As we can observe if we split the dependent variable rating in 2 levels as
Above Average and Below Average
#the pseudo r-square value is very low and the p-value is low as well.
#Also dependent variable rating is not of binomial type and hence the
logistic regression
#model does not fit for our data.
```