Final Assignment 16012993 JunHee Lee

```
#Logistic Regression
# How Teams in EPL and SerieA will move actively in the transfer market
# Predict Transfer based on the data Goals/Loss
dataset = read.csv("soccer.csv") #read dataset-->soccer.csv
str(dataset) #Transfer --> change 'int' to categorical(factor)
```

```
## 'data.frame':    40 obs. of  4 variables:
##  $ Team    : chr  "Liverpool" "ManCity" "Leicster" "Chelsea" ...
##  $ Goals   : int  70 77 59 55 48 44 50 30 43 28 ...
##  $ Loss    : int  21 33 29 41 31 34 41 31 41 36 ...
##  $ Transfer: int  1 1 1 1 1 1 0 0 0 0 ...
```

```
dataset = dataset[2:4] #preprocess the dataset -->[Goals,Loss,Transfer]
head(dataset)
```

```
##   Goals Loss Transfer
## 1    70   21        1
## 2    77   33        1
## 3    59   29        1
## 4    55   41        1
## 5    48   31        1
## 6    44   34        1
```

```
str(dataset) # 'data.frame':    40 obs. of  3 variables:
```

```
## 'data.frame':    40 obs. of  3 variables:
##  $ Goals   : int  70 77 59 55 48 44 50 30 43 28 ...
##  $ Loss    : int  21 33 29 41 31 34 41 31 41 36 ...
##  $ Transfer: int  1 1 1 1 1 1 0 0 0 0 ...
```

```
#Change the target variable(Transfer) to categorical variable(as factor)
dataset$Transfer = factor(dataset$Transfer,levels = c(0,1))
```

```
#install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$Transfer,SplitRatio=0.65) #26Teams for training/14Teams for test
ing.
training_set = subset(dataset,split==TRUE)#Model to predict the transfer market
test_set = subset(dataset,split==FALSE) #Used to validate the model's performance
head(training_set)
```

```
##    Goals Loss Transfer
## 1     70   21        1
## 3     59   29        1
## 6     44   34        1
## 8     30   31        0
## 10    28   36        0
## 11    35   45        0
```

```
head(test_set)
```

```
##    Goals Loss Transfer
## 2     77   33        1
## 4     55   41        1
## 5     48   31        1
## 7     50   41        0
## 9     43   41        0
## 13    29   42        1
```

```
#Feature scaling to make the training/test model. Scale the dataset to normalize data
training_set[,1:2] = scale(training_set[,1:2]) #consecutive values for training set(26 Teams)
training_set[,1:2] #[Goals,Loss]
```

```
##         Goals       Loss
## 1   2.3972790 -1.7743405
## 3   1.5294033 -1.0054596
## 6   0.3459365 -0.5249091
## 8  -0.7586326 -0.8132394
## 10 -0.9164282 -0.3326888
## 11 -0.3641436  0.5323021
## 12 -0.1274503  0.6284123
## 14 -0.1274503  1.3972931
## 15 -0.4430414  0.1478617
## 16 -0.9164282  0.6284123
## 18 -0.8375304  1.0128527
## 21  1.2927099 -1.4860102
## 22  1.7660967 -1.2937899
## 23  1.1349144 -1.1015697
## 25  1.0560166 -0.3326888
## 26  0.2670387 -0.3326888
## 27 -0.2063481 -0.6210192
## 29 -0.6797348 -1.0054596
## 30  0.2670387  0.2439718
## 32  0.4248342  0.6284123
## 33 -0.5219392 -0.2365787
## 34 -0.7586326  0.6284123
## 37 -0.6008370  1.1089628
## 38 -0.3641436  2.3583942
## 39 -1.5476105  0.5323021
## 40 -1.3109171  1.0128527
```

```
test_set[,1:2] = scale(test_set[,1:2]) #consecutive values for test set(14 Teams)
test_set[,1:2] #[Goals,Loss]
```

```
##          Goals       Loss
## 2   1.97276077 -1.1508364
## 4   0.71215845 -0.2168242
## 5   0.31105771 -1.3843394
## 7   0.42565792 -0.2168242
## 9   0.02455719 -0.2168242
## 13 -0.77764429 -0.1000727
## 17 -0.43384365  1.3009455
## 19 -0.37654355  1.8847031
## 20 -1.00684471  1.5344485
## 24  1.97276077 -0.6838303
## 28 -0.60574397 -0.9173333
## 31 -0.26194334  0.1334303
## 35 -1.23604513 -0.5670788
## 36 -0.72034418  0.6004364
```

```
# Fit logistic regression to the training set
classifier = glm(formula = Transfer ~.,family=binomial,data=training_set)
#Transfer --> Dependent(Target)Variable
#formula = Transfer ~. --> use whole independent variables[Y = a + b1X1 + b2X2]
# --> Transfer = a + GoalsX1 + LossX2
```

```
#Predict test set results--> Validation of the training set model
prob_pred = predict(classifier,type='response',newdata = test_set[,1:2])
prob_pred #-->probality to participate in the transfer market
```

```
##         2         4         5         7         9        13        17        19
## 0.8753183 0.7313993 0.5587556 0.6724308 0.5802142 0.3940226 0.5872708 0.6443369
##        20        24        28        31        35        36
## 0.4654654 0.8905969 0.3728603 0.5379621 0.2629837 0.4621019
```

```
# 24(Inter):0.8905969--> Team director Marotta has been actively scouting players to play for
Inter to win the SerieA League and end Juventus's lead
# 2(Liverpool):0.8753183--> Liverpool,the champion of this season will plan bringing in playe
rs to back up current players and to keep building the team
# 5(Chelsea):0.5587556--> After suspension from entering the market, Chelsea is targeting mul
tiple players to once again become the League champion
```

```
y_pred = ifelse(prob_pred >0.5,1,0) # more than 0.5-->goes to 1(active in transfer market)/ l
ess than 0.5--> goes to 0(inactive in the transfer market)
y_pred
```

```
##  2  4  5  7  9 13 17 19 20 24 28 31 35 36
##  1  1  1  1  1  0  1  1  0  1  0  1  0  0
```

```
#Confusion matrix --> compare real and predicted values to assess the model.
cm = table(test_set[,3],y_pred)
cm    #accuracy(3/14), error(11/14) --> tells that making successful signs and contracts in t
ransfer markets is difficult
```

```
##    y_pred
##     0 1
##   0 0 6
##   1 5 3
```

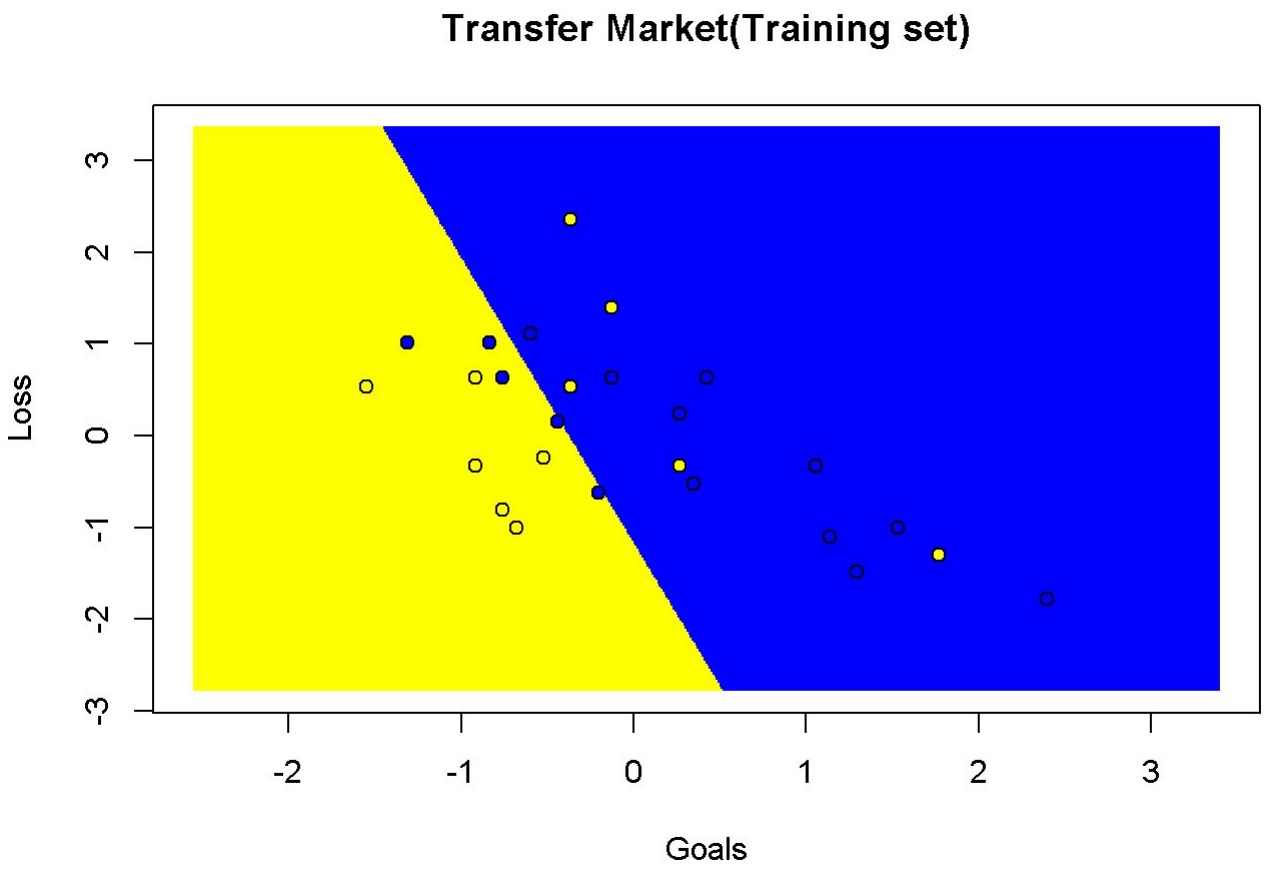
```
(cm1 = table(test_set[,3],y_pred>0.5))
```

```
##
##     FALSE TRUE
##   0     0    6
##   1     5    3
```

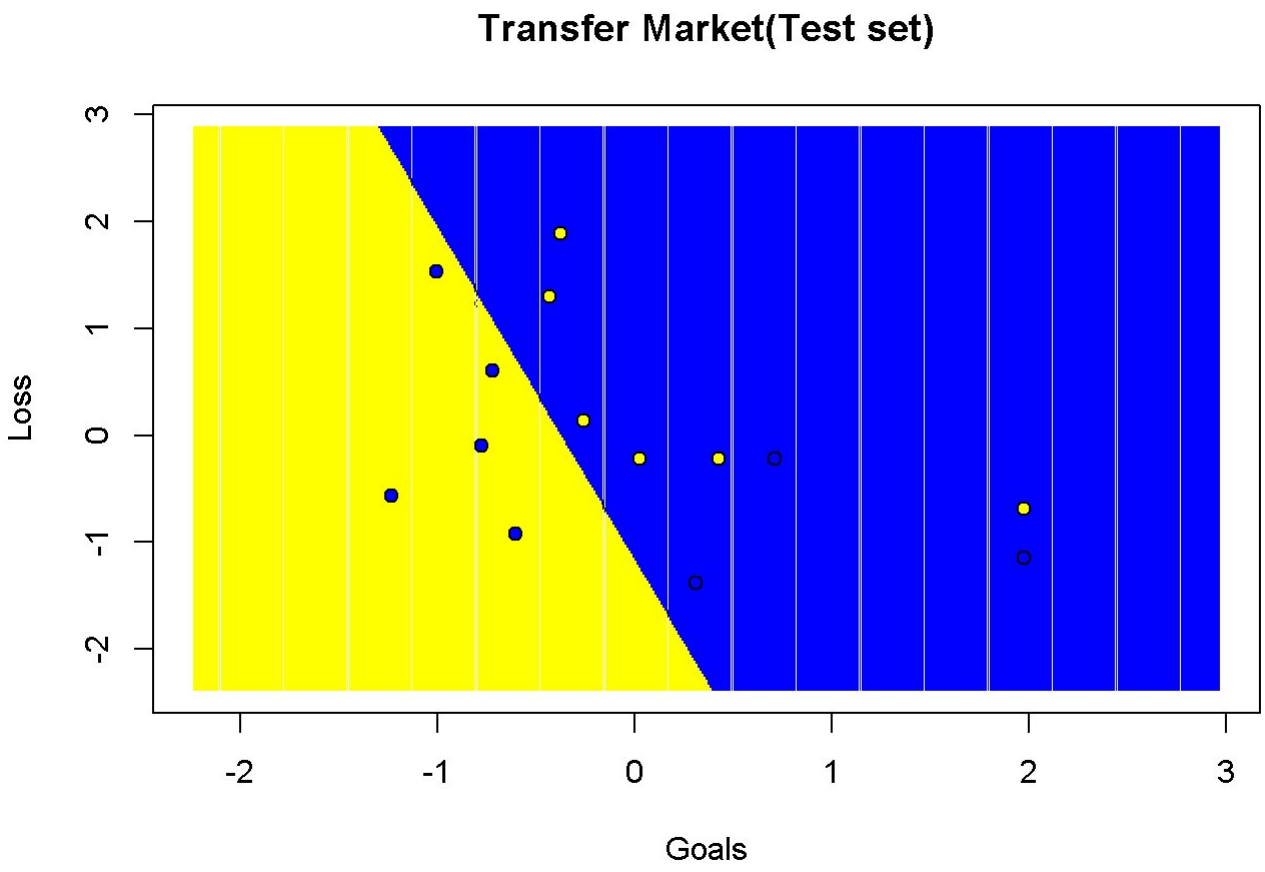
```
test_set #test set to assess the training model.
```

```
##          Goals       Loss Transfer
## 2   1.97276077 -1.1508364        1
## 4   0.71215845 -0.2168242        1
## 5   0.31105771 -1.3843394        1
## 7   0.42565792 -0.2168242        0
## 9   0.02455719 -0.2168242        0
## 13 -0.77764429 -0.1000727        1
## 17 -0.43384365  1.3009455        0
## 19 -0.37654355  1.8847031        0
## 20 -1.00684471  1.5344485        1
## 24  1.97276077 -0.6838303        0
## 28 -0.60574397 -0.9173333        1
## 31 -0.26194334  0.1334303        0
## 35 -1.23604513 -0.5670788        1
## 36 -0.72034418  0.6004364        1
```

```
#Visualization--> Training Set
library(ElemStatLearn)
set = training_set
X1 = seq(min(set[,1]) -1,max(set[,1]) +1,by=0.01)
X2 = seq(min(set[,2]) -1,max(set[,2]) +1,by=0.01)
grid_set = expand.grid(X1,X2)
colnames(grid_set) = c('Goals','Loss')
prob_set = predict(classifier, type = 'response', newdata = grid_set)
y_grid = ifelse(prob_set > 0.5,1,0)
plot(set[,-3],main = 'Transfer Market(Training set)',xlab = 'Goals', ylab = 'Loss',xlim = ran
ge(X1),ylim = range(X2))
contour(X1,X2,matrix(as.numeric(y_grid),length(X1),length(X2)),add=TRUE)
points(grid_set,pch='.',col=ifelse(y_grid==1,'blue','yellow'))
points(set,pch=21,bg=ifelse(set[,3]==1,'blue','yellow'))
```




```
#Visualization --> Test Set
library(ElemStatLearn)
set = test_set
X1 = seq(min(set[,1]) -1,max(set[,1]) +1,by=0.01)
X2 = seq(min(set[,2]) -1,max(set[,2]) +1,by=0.01)
grid_set = expand.grid(X1,X2)
colnames(grid_set) = c('Goals','Loss')
prob_set = predict(classifier, type = 'response', newdata = grid_set)
y_grid = ifelse(prob_set > 0.5,1,0)
plot(set[,-3],main = 'Transfer Market(Test set)',xlab = 'Goals', ylab = 'Loss',xlim = range(X
1),ylim = range(X2))
contour(X1,X2,matrix(as.numeric(y_grid),length(X1),length(X2)),add=TRUE)
points(grid_set,pch='.',col=ifelse(y_grid==1,'blue','yellow'))
points(set,pch=21,bg=ifelse(set[,3]==1,'blue','yellow'))
```




```
# Blue --> active in transfer market
# Yellow --> inactive in transfer market
# Comparing the training and test set logistic regression charts, at first it seems that the
teams with the higher rank of each league will be active in the transfer market. This is alwa
ys just expectable because good teams put more money in the market to upgrade their squad. Bu
t there's a big difference in the test set. This shows the complexity of deals among teams an
d many other reasons of the difficulty such as agency,team money, the director's policy, play
ers' desire etc.

# In transfer markets in soccer leagues, many untrustable issues occur and there many volatil
ity of signing new contracts between teams in case of buying and selling players. For exampl
e, it was expected that Ivan Persic of Inter will be sold to Bayern Munchen, but Bayern sudde
nly refused to buy. And there was a sudden deal with Achraf Hakimi between Dortmund and Inte
r. So, linear logistic regression shows some outliers according to these issues. And the trai
ning set and training set also shows some differences according to these issues.
```