WEEK 9 :CLASSIFICATION MODEL

PROBLEM DEFINATION:
g. Install relevant package for classification.

SOURCE CODE:
```
install.packages("rpart.plot")
install.packages("tree")
install.packages("ISLR")
install.packages("rattle")

library(tree)
library(ISLR)
library(rpart.plot)
library(rattle)
```

PROBLEM DEFINATION:
h. Choose classifier for classification problem.
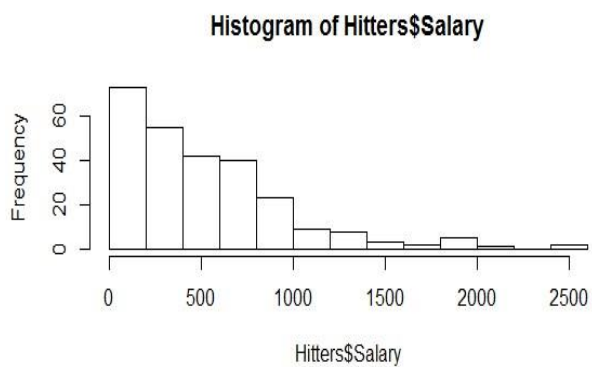Evaluate the performance of classifier**.**
SOURCE CODE:
```
attach(Hitters)
View(Hitters)
# Remove NA data
Hitters<-na.omit(Hitters)

# log transform Salary to make it a bit more normally distributed
hist(Hitters$Salary)

Hitters$Salary <- log(Hitters$Salary)
hist(Hitters$Salary)
```

output:

Histogram of Hitters$Salary

SOURCE CODE:
```
> tree.fit <- tree(Salary~Hits+Years, data=Hitters)
> summary(tree.fit)
```

```
Regression tree:
tree(formula = Salary ~ Hits + Years, data = Hitters)
Number of terminal nodes:  8
Residual mean deviance:  101200 = 25820000 / 255
Distribution of residuals:
   Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
-1238.00  -157.50  -38.84    0.00   76.83  1511.00
```

```
plot(tree.fit, uniform=TRUE,margin=0.2)
text(tree.fit, use.n=TRUE, all=TRUE, cex=.8)
#plot(tree.fit)
>split <- createDataPartition(y=Hitters$Salary, p=0.5, list=FALSE)
> train <- Hitters[split,]
> test <- Hitters[-split,]
#Create tree model
> trees <- tree(Salary~., train)
> plot(trees)
> text(trees, pretty=0)
```
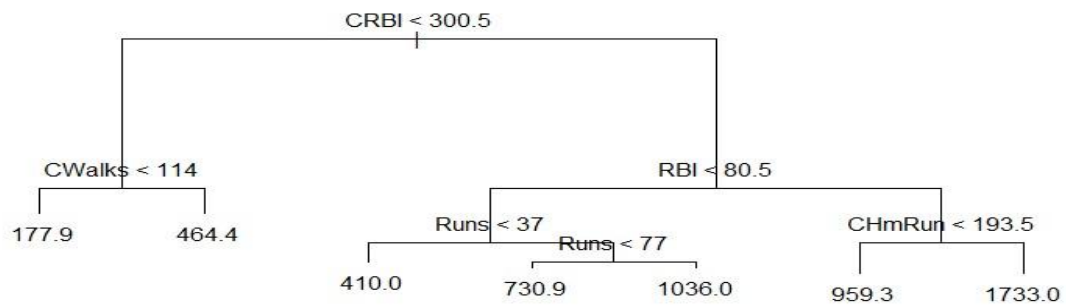
```
# Cross validate to see whether pruning the tree will improve
 Performance
```

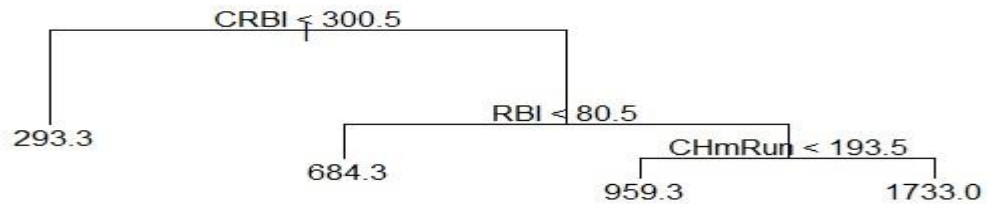**OUTPUT:**

SOURCE CODE:
#Cross validate to see whether pruning the tree will improve performance
> cv.trees <- cv.tree(trees)
> plot(cv.trees)
> prune.trees <- prune.tree(trees, best=4)
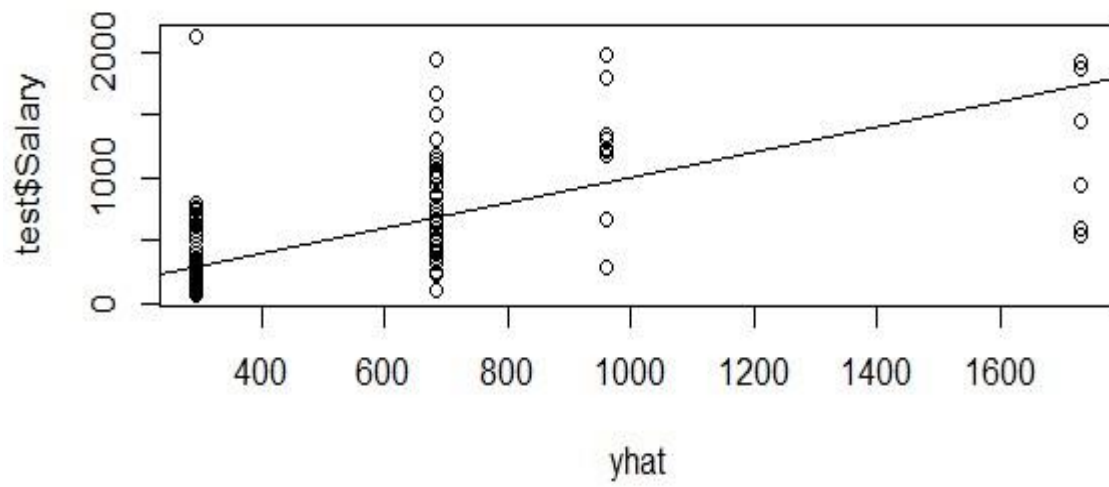> plot(prune.trees)
> text(prune.trees, pretty=0)

OUTPUT:

SOURCE CODE:
> yhat <- predict(prune.trees, test)
> plot(yhat, test$Salary)
> abline(0,1
[1] 150179.7
> mean((yhat - test$Salary)^2)
[1] 150179.7

OUTPUT:

```
> mean((yhat - test$Salary)^2)
[1] 150179.7
```