Machine Learning Models

1. Load Libraries.

```
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:randomForest':
##
##     margin
```

```
library(rpart)
```

2. Load test data.

```
setwd("c:/Ryerson University/Semester 4/ProjectCode")
loc<-getwd()
traindata <- read.csv(file="traindata.csv",header=TRUE,sep=",")
testdata1 <- read.csv(file="testdata1.csv",header=TRUE,sep=",")
dim(testdata1)
```

```
## [1] 7547   15
```

```
#####################################################################
#                        LOGISTIC REGRESION
# Regression coefficients represent the mean change in the response variable for one unit of change # i
#####################################################################
m1 <- glm(income ~ age+ workclass+ education+marital.status+ occupation+relationship+race+ sex +hours.pe
summary(m1)
```

```
##
## Call:
## glm(formula = income ~ age + workclass + education + marital.status +
##     occupation + relationship + race + sex + hours.per.week,
##     family = binomial("logit"), data = traindata)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.7320  -0.5767  -0.2212  -0.0012   3.3378
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -7.589216   0.464650 -16.333  < 2e-16
## age                              0.029287   0.001852  15.816  < 2e-16
## workclassOther                   0.704063   0.116072   6.066 1.31e-09
## workclassPrivate                 0.257853   0.067809   3.803 0.000143
## workclassSelf-Employed           0.004985   0.083539   0.060 0.952414
## workclassWithout-pay           -12.510686 249.291349  -0.050 0.959975
## education11th                    0.083280   0.234625   0.355 0.722626
## education12th                    0.614487   0.285902   2.149 0.031611
## education1st-4th                -0.956547   0.549775  -1.740 0.081879
## education5th-6th                -0.784810   0.405625  -1.935 0.053013
## education7th-8th                -0.619452   0.258370  -2.398 0.016506
## education9th                    -0.254599   0.277626  -0.917 0.359113
## educationAssoc-acdm              1.240166   0.190374   6.514 7.30e-11
## educationAssoc-voc               1.276327   0.181804   7.020 2.21e-12
## educationBachelors               1.921347   0.168791  11.383  < 2e-16
## educationDoctorate               2.872910   0.236375  12.154  < 2e-16
## educationHS-grad                 0.770195   0.163647   4.706 2.52e-06
## educationMasters                 2.324071   0.180927  12.845  < 2e-16
## educationPreschool             -11.554436 149.080039  -0.078 0.938222
## educationProf-school             2.909688   0.215362  13.511  < 2e-16
## educationSome-college            1.069891   0.166360   6.431 1.27e-10
## marital.statusMarried-AF-spouse  2.089808   0.632998   3.301 0.000962
## marital.statusMarried-civ-spouse 1.844631   0.316700   5.825 5.73e-09
## marital.statusMarried-spouse-absent -0.146120   0.252539  -0.579 0.562857
## marital.statusNever-married     -0.400126   0.094195  -4.248 2.16e-05
## marital.statusSeparated         -0.002376   0.171650  -0.014 0.988956
## marital.statusWidowed            0.162155   0.164934   0.983 0.325534
## occupationArmed-Forces          -0.884259   1.286578  -0.687 0.491897
## occupationCraft-repair           0.025056   0.087571   0.286 0.774786
## occupationExec-managerial        0.837831   0.084433   9.923  < 2e-16
## occupationFarming-fishing       -1.055271   0.148750  -7.094 1.30e-12
## occupationHandlers-cleaners     -0.736631   0.156706  -4.701 2.59e-06
## occupationMachine-op-inspct     -0.333986   0.112855  -2.959 0.003082
## occupationOther-service         -0.934145   0.130186  -7.175 7.20e-13
## occupationPriv-house-serv       -2.324552   1.155617  -2.012 0.044270
## occupationProf-specialty         0.528958   0.089536   5.908 3.47e-09
## occupationProtective-serv        0.440954   0.137817   3.200 0.001376
## occupationSales                  0.317584   0.089717   3.540 0.000400
## occupationTech-support           0.521752   0.123270   4.233 2.31e-05
## occupationTransport-moving      -0.122669   0.109070  -1.125 0.260724
## relationshipNot-in-family        0.245985   0.313494   0.785 0.432655
## relationshipOther-relative      -0.646159   0.288232  -2.242 0.024974
## relationshipOwn-child           -0.887215   0.313869  -2.827 0.004703
## relationshipUnmarried            0.112798   0.327938   0.344 0.730876
## relationshipWife                 1.286684   0.113770  11.310  < 2e-16
## raceAsian-Pac-Islander           0.336590   0.269814   1.247 0.212217
## raceBlack                        0.389583   0.255037   1.528 0.126623
## raceOther                       -0.280137   0.393981  -0.711 0.477059
## raceWhite                        0.528277   0.243263   2.172 0.029883
```

```
## sexMale                                0.835270   0.085317   9.790  < 2e-16
## hours.per.week                         0.031585   0.001857  17.007  < 2e-16
##
## (Intercept)                           ***
## age                                   ***
## workclassOther                        ***
## workclassPrivate                      ***
## workclassSelf-Employed
## workclassWithout-pay
## education11th
## education12th                         *
## education1st-4th                      .
## education5th-6th                      .
## education7th-8th                      *
## education9th
## educationAssoc-acdm                   ***
## educationAssoc-voc                    ***
## educationBachelors                    ***
## educationDoctorate                    ***
## educationHS-grad                      ***
## educationMasters                      ***
## educationPreschool
## educationProf-school                  ***
## educationSome-college                 ***
## marital.statusMarried-AF-spouse       ***
## marital.statusMarried-civ-spouse      ***
## marital.statusMarried-spouse-absent
## marital.statusNever-married           ***
## marital.statusSeparated
## marital.statusWidowed
## occupationArmed-Forces
## occupationCraft-repair
## occupationExec-managerial             ***
## occupationFarming-fishing             ***
## occupationHandlers-cleaners           ***
## occupationMachine-op-inspct           **
## occupationOther-service               ***
## occupationPriv-house-serv             *
## occupationProf-specialty              ***
## occupationProtective-serv             **
## occupationSales                       ***
## occupationTech-support                ***
## occupationTransport-moving
## relationshipNot-in-family
## relationshipOther-relative            *
## relationshipOwn-child                 **
## relationshipUnmarried
## relationshipWife                      ***
## raceAsian-Pac-Islander
## raceBlack
## raceOther
## raceWhite                             *
## sexMale                               ***
## hours.per.week                        ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25389  on 22614  degrees of freedom
## Residual deviance: 16218  on 22564  degrees of freedom
## AIC: 16320
##
## Number of Fisher Scoring iterations: 13
```

```
prob <- predict(m1, testdata1, type = 'response')
prediction <- predict(m1,testdata1,type='response')

#############################################################################
# P values shows that Age ,workclass, education, marital status, occupation,
# race, sex, hours per week  are the significant attributes.
#############################################################################
pred <- rep('<=50K', length(prob))
pred[prob>=.5] <- '>50K'
tb <- table(pred, testdata1$income)
tb
```

```
##
## pred    <=50K >50K
##   <=50K  5190  792
##   >50K    482 1083
```

```
# Confusion matrix shows that it has an Accuracy of 83.11%
# misclasification 16.8%.
```

DECISION TREE

```
Dtree<- rpart(income~ age+ workclass+ education+marital.status+ occupation+relationship+race+ sex +hours
Dtree.pred.prob <- predict(Dtree, newdata = testdata1, type = 'prob')
Dtree.pred <- predict(Dtree, newdata = testdata1, type = 'class')
confusionMatrix(testdata1$income,Dtree.pred)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  5218  454
##      >50K    814 1061
##
##                Accuracy : 0.832
##                  95% CI : (0.8234, 0.8404)
##     No Information Rate : 0.7993
##     P-Value [Acc > NIR] : 2.294e-13
##
##                   Kappa : 0.5192
##  Mcnemar's Test P-Value : < 2.2e-16
##
```

```
##             Sensitivity : 0.8651
##             Specificity : 0.7003
##          Pos Pred Value : 0.9200
##          Neg Pred Value : 0.5659
##              Prevalence : 0.7993
##          Detection Rate : 0.6914
##    Detection Prevalence : 0.7516
##       Balanced Accuracy : 0.7827
##
##        'Positive' Class : <=50K
##
```

```
# Confusion matrix shows that it has an Accuray of 83.2%
```

RANDOM FOREST

```
library(randomForest)
levels(testdata1$workclass) <- levels(traindata$workclass)
rforest <- randomForest(income ~ age+ workclass+ education+marital.status+occupation+relationship+race+
rforest.pred.prob <- predict(rforest, newdata = testdata1, type = 'prob')
rforest.pred <- predict(rforest, newdata = testdata1, type = 'class')
# confusion matrix
tb3 <- table(rforest.pred, testdata1$income)
tb3
```

```
##
## rforest.pred <=50K >50K
##        <=50K  5136  698
##        >50K    536 1177
```

```
confusionMatrix(testdata1$income,rforest.pred)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  5136  536
##      >50K    698 1177
##
##                Accuracy : 0.8365
##                  95% CI : (0.828, 0.8448)
##     No Information Rate : 0.773
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5491
##  Mcnemar's Test P-Value : 4.579e-06
##
##             Sensitivity : 0.8804
##             Specificity : 0.6871
##          Pos Pred Value : 0.9055
##          Neg Pred Value : 0.6277
##              Prevalence : 0.7730
##          Detection Rate : 0.6805
```

```
##    Detection Prevalence : 0.7516
##       Balanced Accuracy : 0.7837
##
##        'Positive' Class : <=50K
##
```

```
## LINEAR REGRESION
pr  <- prediction(prob,testdata1$income)
perf <- performance(pr,measure="tpr", x.measure="fpr")
DtFrameReg <- data.frame(FP=perf@x.values[[1]],TP=perf@y.values[[1]])
aucRegresion <- performance(pr,measure='auc')@y.values[[1]]
aucRegresion
```

```
## [1] 0.8859387
```

```
###DECISION TREE
prtree <- prediction(Dtree.pred.prob[,2],testdata1$income)
perftree  <- performance(prtree,measure="tpr",x.measure="fpr")
DTFrametree <- data.frame(FP=perftree@x.values[[1]],TP=perftree@y.values[[1]])
auctree <- performance(prtree, measure='auc')@y.values[[1]]
auctree
```

```
## [1] 0.8529889
```

```
###RANDOM FOREST
prRForest <- prediction(rforest.pred.prob[,2],testdata1$income)
perfRForest  <- performance(prRForest,measure="tpr",x.measure="fpr")
DTFrameRForest <- data.frame(FP=perfRForest@x.values[[1]],TP=perfRForest@y.values[[1]])
auctree <- performance(prRForest, measure='auc')@y.values[[1]]
auctree
```

```
## [1] 0.8755331
```