Prediction Challenge 04

Jack Lin

Model 1: Rpart

R code:

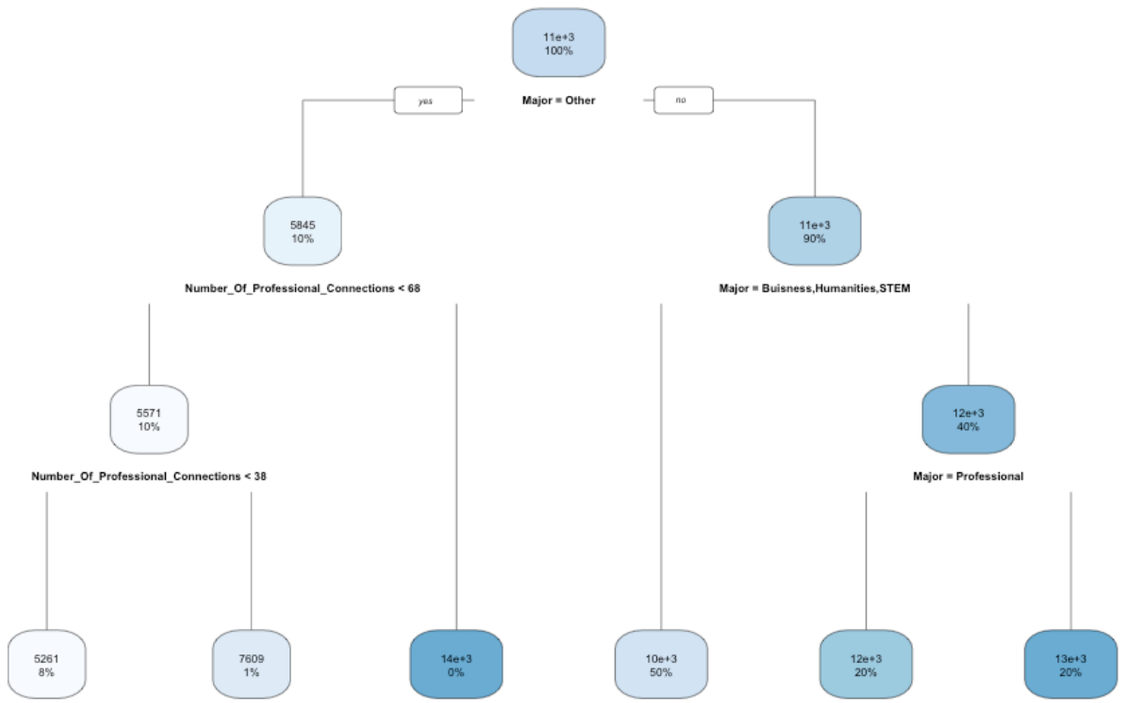
```
> Earnings_Numeric_Train <- read.csv("~/Documents/Rutgers/Data 101/Assignments/Prediction Challenge 4/Earnings_Numeric_Train.csv", stringsAsFactors=FALSE)
>   View(Earnings_Numeric_Train)
> Earnings_Test_Students._2_ <- read.csv("~/Documents/Rutgers/Data 101/Assignments/Prediction Challenge 4/Earnings_Test_Students _2_.csv",
stringsAsFactors=FALSE)
>   View(Earnings_Test_Students._2_)
> library(rpart)
> library(rpart.plot)
> earning.tree <- rpart(Earnings~Major+Number_Of_Professional_Connections,data=Earnings_Numeric_Train,control=rpart.control(minsplit=50))
> rpart.plot(earning.tree)
> predict.rpart <- predict(earning.tree,newdata=Earnings_Numeric_Train)
> mse.rpart <- mean((predict.rpart - Earnings_Numeric_Train$Earnings)^2)
> mse.rpart
[1] 93608.69
> rmse.rpart <- sqrt(mean((predict.rpart - Earnings_Numeric_Train$Earnings)^2))
> rmse.rpart
[1] 305.9554
> CrossValidation::cross_validate(Earnings_Numeric_Train,ear
[[1]]
   accuracy_subset accuracy_all
1         69331.81     69331.81
2         56368.69     56368.69
3         70697.80     70697.80
4        100305.17    100305.17
5         66348.98     66348.98
6        132632.21    132632.21
7         96537.19     96537.19
8        204610.99    204610.99
9         69518.09     69518.09
10       144815.52    102137.20

[[2]]
[[2]]$average_accuracy_subset
[1] 101116.6

[[2]]$average_accuracy_all
[1] 96848.81

[[2]]$variance_accuracy_subset
[1] 2197861419

[[2]]$variance_accuracy_all
[1] 1965562147
```



Model 2: LM

R code:

```
> earning.lm <- lm(Earnings~.,data=Earnings_Numeric_Train)
> earning.lm

Call:
lm(formula = Earnings ~ ., data = Earnings_Numeric_Train)

Coefficients:
                (Intercept)
                  9970.9468
                        GPA
                    16.8303
Number_Of_Professional_Connections
                     8.5327
                MajorHumanities
                   252.2320
                   MajorOther
                 -4155.4819
              MajorProfessional
                  1748.5176
                   MajorSTEM
                  -252.9820
              MajorVocational
                  3246.4067
              Graduation_Year
                    -0.1669
                      Height
                    -0.1237
             Number_Of_Credits
                     1.2719
       Number_Of_Parking_Tickets
                    -0.3780
> predict.lm <- predict(earning.lm,newdata=Earnings_Numeric_Train)
> mse.lm <- mean((predict.lm - Earnings_Numeric_Train$Earnings)^2)
> mse.lm
[1] 318142.3
> rmse.lm <- sqrt(mean((predict.lm - Earnings_Numeric_Train$Earnings)^2))
> rmse.lm
[1] 564.041
> CVLinearModel::cross_validatelm(Earnings_Numeric_Train,earning.lm,10,0.8)
   accuracy_subset accuracy_all
1         309251.5     309251.5
2         325203.6     325203.6
3         299646.8     299646.8
4         339590.2     339590.2
5         139258.3     139258.3
6         417077.9     417077.9
7         280572.9     280572.9
8         177075.2     177075.2
9         374496.2     374496.2
10        252026.5     252026.5
```

Model 3: SVM

R code:

```
> library(e1071)
> earning.svm <- svm(Earnings~.,data=Earnings_Numeric_Train)
> earning.svm

Call:
svm(formula = Earnings ~ ., data = Earnings_Numeric_Train)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.08333333
    epsilon:  0.1


Number of Support Vectors:  1390
> predict.svm <- predict(earning.svm,newdata=Earnings_Numeric_Train)
> mse.svm <- mean((predict.svm - Earnings_Numeric_Train$Earnings)^2)
> mse.svm
[1] 53711.97
> rmse.svm <- sqrt(mean((predict.svm - Earnings_Numeric_Train$Earnings)^2))
> rmse.svm
[1] 231.7584
> CrossValidation::cross_validate(Earnings_Numeric_Train,earning.svm,10,0.8,method="anova")
[[1]]
   accuracy_subset accuracy_all
1        101964.16    101964.16
2         97575.41     97575.41
3        140480.33    140480.33
4        137965.48    137965.48
5        249281.82    249281.82
6        103522.83    103522.83
7         63106.38     63106.38
8        195723.14    195723.14
9         54959.98     54959.98
10        75418.02     75418.02

[[2]]
[[2]]$average_accuracy_subset
[1] 121999.8

[[2]]$average_accuracy_all
[1] 121999.8

[[2]]$variance_accuracy_subset
[1] 3744920834

[[2]]$variance_accuracy_all
[1] 3744920834
```

Implementing my SVM model

R code:
> Earnings_Test_Students._2_ <- read.csv("~/Documents/Rutgers/Data 101/Assignments/Prediction Challenge 4/Earnings_Test_Students _2_.csv", stringsAsFactors=FALSE)
> View(Earnings_Test_Students._2_)
> Earnings_Test_Students._2_$Earnings <- predict(earning.svm,newdata=Earnings_Test_Students._2_)
> write.csv(Earnings_Test_Students._2_,”YuHonLinSubmission04.csv”)

I manually deleted columns I don’t need on MS Excel.

Results: RMSE=194.25282 on Kaggle public leaderboard.