MSE Value generated from rpart based Prediction Model:

Mse = 8439.938

This value was generated using this code:

library(rpart)

library(rpart.plot)

# Read the data

train <- read.csv("D:/Uni Resources/Data 101/Project Component 1/EarningsTrain.csv")

set.seed(123)

# Split the data into training and test sets

TrainIndex <- sample(nrow(train), 0.7 \* nrow(train))

TrainData <- train[TrainIndex, ]

TestData <- train[-TrainIndex, ]

num\_vars <- c("Number\_Of\_Professional\_Connections", "GPA", "Earnings",

"Graduation\_Year", "Height", "Number\_Of\_Credits", "Number\_Of\_Parking\_Tickets")

get\_out <- function(data, num\_vars, threshold = 2.9) {

for (var in num\_vars) {

z\_scores <- abs(scale(data[[var]]))

data <- data[z\_scores < threshold, ]

}

return(data)

}

TrainData <- get\_out(TrainData, num\_vars)

TrainData$pred1 <- (TrainData$GPA \* TrainData$Number\_Of\_Professional\_Connections^2)

TrainData$pred3 <- (TrainData$Number\_Of\_Professional\_Connections \* TrainData$Graduation\_Year)

minBuck <-16

minS <- 15

cup <- 0.0001

dep <- 4

# Build regression tree models all majors

tree1 <- rpart(formula = Earnings ~ (Number\_Of\_Professional\_Connections^2) + pred3,

data = TrainData[TrainData$Major == "Other", ],

method = "anova",

control = rpart.control(minsplit = minS, minbucket = minBuck, cp = cup, maxdepth = dep))

tree2 <- rpart(formula = Earnings ~ .,

data = TrainData[TrainData$Major == "STEM", ],

method = "anova",

control = rpart.control(minsplit = minS, minbucket = minBuck, cp = cup, maxdepth = dep))

tree3 <- rpart(formula = Earnings ~ .,

data = TrainData[TrainData$Major == "Vocational", ],

method = "anova",

control = rpart.control(minsplit = minS, minbucket = minBuck, cp = cup, maxdepth = dep))

tree4 <- rpart(formula = Earnings ~ .,

data = TrainData[TrainData$Major == "Buisness", ],

method = "anova",

control = rpart.control(minsplit = minS, minbucket = minBuck, cp = cup, maxdepth = dep))

tree5 <- rpart(formula = Earnings ~ .,

data = TrainData[TrainData$Major == "Humanities", ],

method = "anova",

control = rpart.control(minsplit = minS, minbucket = minBuck, cp = cup, maxdepth = dep))

tree6 <- rpart(formula = Earnings ~ .,

data = TrainData[TrainData$Major == "Professional", ],

method = "anova",

control = rpart.control(minsplit = minS, minbucket = minBuck, cp = cup, maxdepth = dep))

# Make predictions for all majors separately

TrainData$pred.Earnings[TrainData$Major == "Other"] <- predict(tree1, newdata = TrainData[TrainData$Major == "Other", ])

TrainData$pred.Earnings[TrainData$Major == "STEM"] <- predict(tree2, newdata = TrainData[TrainData$Major == "STEM", ])

TrainData$pred.Earnings[TrainData$Major == "Vocational"] <- predict(tree3, newdata = TrainData[TrainData$Major == "Vocational",])

TrainData$pred.Earnings[TrainData$Major == "Buisness"] <- predict(tree4, newdata = TrainData[TrainData$Major == "Buisness", ])

TrainData$pred.Earnings[TrainData$Major == "Humanities"] <- predict(tree5, newdata = TrainData[TrainData$Major == "Humanities", ])

TrainData$pred.Earnings[TrainData$Major == "Professional"] <- predict(tree6, newdata = TrainData[TrainData$Major == "Professional", ])

# Calculate MSE for combined predictions

mse <- mean((TrainData$Earnings - TrainData$pred.Earnings)^2)

print(mse)