library(ggplot2)

library(ggthemes)

library(dplyr)

library(caTools)

df <- read.csv('student-mat.csv',sep=';')

head(df)

# Set a random see so your "random" results are the same as this notebook

set.seed(101)

# Split up the sample, basically randomly assigns a booleans to a new column "sample"

sample <- sample.split(df$age, SplitRatio = 0.70) # SplitRatio = percent of sample==TRUE

#nrow(df)

#sample

# Training Data

train = subset(df, sample == TRUE)

# Testing Data

test = subset(df, sample == FALSE)

model <- lm(G3 ~ .,train)

summary(model)

G3.predictions <- predict(model,test)

results <- cbind(G3.predictions,test$G3)

colnames(results) <- c('pred','real')

results <- as.data.frame(results)

results

to\_zero <- function(x){

if (x < 0){

return(0)

}else{

return(x)

}

}

results$pred <- sapply(results$pred,to\_zero)

mse <- mean((results$real-results$pred)^2)

print(mse)

results

SSE <- sum((results$pred - results$real)^2)

SST <- sum( (mean(df$G3) - results$real)^2)

R2 <- 1 - SSE/SST

R2

#newdata <- data.frame(model\_year=70, weight=3691, origin=1, horsepower = 200)

#predict(model, newdata, interval = "confidence")

#help("predict")

model2 <- lm(G3 ~ famrel+absences+G2,train)

summary(model2)

G3.predictions2 <- predict(model2,test)

results2 <- cbind(G3.predictions2,test$G3)

colnames(results2) <- c('pred','real')

results2

results2 <- as.data.frame(results2)

results2$pred <- sapply(results2$pred,to\_zero)

mse <- mean((results2$real-results2$pred)^2)

print(mse)

results2

SSE2 <- sum((results2$pred - results2$real)^2)

SST2 <- sum( (mean(df$G3) - results2$real)^2)

R3 <- 1 - SSE2/SST2

R3

newdata <- data.frame(famrel=12, absences=10, G2=20)

newdata <- data.frame(famrel=12)#, absences=10, G2=20)

predict(model2, newdata, interval = "confidence")