

dataset Prediction

Predicting Car Origin using Linear Regression



July 18, 2024

**1. Loading Dataset (Replace with your own file path)**

setwd("C:/Users/sikan/Downloads")

csvdata <- read.csv("Auto 2.csv", stringsAsFactors = FALSE)

**2. Calculating horsepower mean & replacing missing values in ‘horsepower’ with mean**

csvdata$horsepower[csvdata$horsepower == '?'] <- NA

csvdata$horsepower <- as.numeric(csvdata$horsepower)

mean\_horsepower <- round(mean(csvdata$horsepower, na.rm = TRUE))

csvdata$horsepower[is.na(csvdata$horsepower)] <- mean\_horsepower

**3. Multiple Linear Regression Model with mpg without including ‘name’**

model <- lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + origin, data = csvdata)

model\_summary <- summary(model)

**4. Identifying significant predictors**

predictors <- model\_summary$coefficients[model\_summary$coefficients[, "Pr(>|t|)"] < 0.05, ]

print(predictors)

**5. Check coefficient for the year variable**

coefficientfortheyear <- model\_summary$coefficients["year", ]

print(coefficientfortheyear)

**6. Making linear regression equation for the model**

linear\_regression\_equation <- paste("mpg =", round(coef(model)[1], 2), "+", paste(paste(names(coef(model))[-1], round(coef(model)[-1], 2), sep = "\*"), collapse = " + "))

print(linear\_regression\_equation)

**7. Predicted Value for the mpg of the car**

**7.1. Creating new dataframe for Prediction**

new\_car <- data.frame(cylinders = 4, displacement = 100, horsepower = 50, weight = 3000, acceleration = 15, year = 75, origin = 1)

**7.2. Predicting mpg for the new car**

predicted\_mpg <- predict(model, newdata = new\_car)

print(predicted\_mpg)

**8. Fitting C5.0 tree to the training data & Defining training and testing set**

**8.1. Install and load necessary library for C5.0 tree**

if (!require(C50)) {

install.packages("C50")

library(C50)

}

**8.2. Randomly sample 300 rows for training set**

set.seed(123)

train\_indices <- sample(nrow(csvdata), 300)

train\_data <- csvdata[train\_indices, ]

**8.3. Use remaining rows for testing set**

test\_data <- csvdata[-train\_indices, ]

**8.3. Fitting C5.0 tree model**

train\_data$origin <- factor(train\_data$origin)

tree\_model <- C5.0(origin ~ cylinders + displacement + horsepower + weight + acceleration + year, data = train\_data)

**9. Showing Tree Size**

print(tree\_model$size)

**10. Drawing Confusion matrix & Calculating error rate**

**10.1. creating a confusion matrix**

predictions <- predict(tree\_model, newdata = test\_data)

conf\_matrix <- table(test\_data$origin, predictions)

print(conf\_matrix)

**10.2. Calculating the error rate**

error\_rate <- 1 - sum(diag(conf\_matrix)) / sum(conf\_matrix)

print(paste("Error Rate:", round(error\_rate, 4)))

**11. Predicting value of origin for the new car using the decision tree**

new\_car <- data.frame(cylinders = 4, displacement = 100, horsepower = 50, weight = 3000, acceleration = 15, year = 75)

predicted\_origin <- predict(tree\_model, newdata = new\_car)

print(predicted\_origin)

savehistory("script.Rhistory")

**Create Description.txt file**

description <- paste(

"1. Loaded the dataset from 'Auto 2.csv'.",

"\n2. Missing values in horsepower were replaced with the mean (rounded to the nearest integer).",

"\n3. Performed multiple linear regression with mpg as the response variable and all other variables except name as predictors.",

"\n4. Predictors with statistically significant relationships (p-value < 0.05):",

paste(names(predictors), collapse = ", "),

"\n5. The coefficient for the year variable suggests that for each one-year increase, the mpg increases by approximately 0.75 units (positive relationship).",

"\n6. Linear regression equation:\n", linear\_regression\_equation,

"\n7. Predicted mpg for the specified car features:\n", round(predicted\_mpg, 2),

"\n8. Fitted a C5.0 decision tree to the training data.",

"\n9. Decision tree size: ", tree\_model$size,

"\n10. Confusion matrix and error rate:\n",

paste(capture.output(print(conf\_matrix)), collapse = "\n"),

"\nError Rate: ", round(error\_rate, 4),

"\n11. Predicted origin for the specified car features: ", predicted\_origin,

sep = "\n"

)

writeLines(description, "C:/Users/sikan/Downloads/description.txt")