Technical Report

Duc Nguyen and Linh Vu

2024-05-10

First note: Below is all the code chunk for data wrangling so that we can run and present the plots later. However, the UCI machine learning website is down, so we cannot load our data. Commenting out all the code lines would be very tedious, so we decided to delete the '}' at the beginning of the code chunk, the downside of which is that our current version of the pdf is very messy. Thank you for your understanding.

"'{r, include=FALSE library(readr) student_data <- read_csv('https://archive.ics.uci.edu/static/public/697/data.csv')

library(dplyr) student_data <- mutate(student_data, `Marital Status` = as.factor(`Marital Status`)) student_data <- mutate(student_data, `Displaced` = as.factor(`Displaced`)) student_data <- mutate(student_data, `Daytime/evening attendance` = as.factor(`Daytime/evening attendance`)) student_data <- mutate(student_data, `Educational special needs` = as.factor(`Educational special needs`)) student_data <- mutate(student_data, `Tuition fees up to date` = as.factor(`Tuition fees up to date`)) student_data <- mutate(student_data, `Gender` = as.factor(`Gender`)) student_data <- mutate(student_data, `Scholarship holder` = as.factor(`Scholarship holder`)) student_data <- mutate(student_data, `Mother's qualification` = as.factor(`Mother's qualification`)) student_data <- mutate(student_data, `Father's qualification` = as.factor(`Father's qualification`)) student_data <- mutate(student_data, `Previous qualification` = as.factor(`Previous qualification`))

names(student_data)[names(student_data) == "Course"] <- "Course_Enrolled" names(student_data)[names(student_data == "Nacionality"] <- "Nationality"

student_data <- mutate(student_data, `Nationality` = as.factor(`Nationality`)) student_data <- mutate(student_data, `Course_Enrolled` = as.factor(`Course_Enrolled`)) student_data <- mutate(student_data, `Mother's occupation` = as.factor(`Mother's occupation`)) student_data <- mutate(student_data, `Father's occupation` = as.factor(`Father's occupation`)) student_data <- mutate(student_data, `Debtor` = as.factor(`Debtor`)) student_data <- mutate(student_data, `International` = as.factor(`International`))

library(forcats)

student_data$Mother's qualification$ <- -factor(student_data`Mother's qualification`, levels = c(1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 14, 18, 19, 22, 26, 27, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44))

Collapse the levels into broader categories

student_data <- mutate(student_data, `Mother's qualification` = fct_collapse(`Mother's qualification`, Basic_Education = c("19", "26", "27", "37", "38"), Secondary_Education = c("1", "9", "12", "14", "18", "29", "30", "10", "11"), Higher_Education = c("2", "3", "4", "5", "6", "40", "41", "42", "43", "44"), Professional_Technical = c("22", "39"), Unknown_None = c("34", "35", "36", "31", "33")))

levels(student_data$`Mother's qualification`)

student_data‘$Father's qualification‘ <- factor(student_data$Father's qualification, levels = c(1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 14, 18, 19, 22, 26, 27, 29, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44))

Collapse the levels into broader categories

student_data <- mutate(student_data, Father's qualification = fct_collapse(Father's qualification, Basic_Education = c("19", "26", "27", "37", "38"), Secondary_Education = c("1", "9", "12", "14", "18", "29", "30", "10", "11"), Higher_Education = c("2", "3", "4", "5", "6", "40", "41", "42", "43", "44"), Professional_Technical = c("22", "39"), Unknown_None = c("34", "35", "36", "31", "33")))

levels(student_data$Father's qualification)

student_data‘$Previous qualification‘ <- factor(student_data$Previous qualification, levels = c(1, 2, 3, 4, 5, 6, 9, 10, 12, 14, 15, 19, 38, 39, 40, 42, 43))

Collapse the factor levels into broader categories using forcats::fct_collapse

student_data <- mutate(student_data, Previous qualification = fct_collapse(Previous qualification, Basic_Education = c("19", "38"), Secondary_Education = c("1", "9", "10", "12", "14", "15"), Higher_Education = c("2", "3", "4", "5", "6", "40", "43"), Professional_Technical = c("39", "42")))

Print the new levels to verify the changes

levels(student_data$Previous qualification)

student_data‘$Mother's occupation‘ <- factor(student_data$Mother's occupation, levels = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 90, 99, 122, 123, 125, 131, 132, 134, 141, 143, 144, 151, 152, 153, 171, 173, 175, 191, 192, 193, 194))

Collapse the levels into broader categories

student_data <- mutate(student_data, Mother's occupation = fct_collapse(Mother's occupation, Student = "0", High_Level_Professionals = c("1", "2", "122", "123", "125"), Intermediate_Professionals = c("3", "131", "132", "134"), Administrative_Staff = c("4", "141", "143", "144"), Service_Workers = c("5", "151", "152", "153", "191"), Skilled_Workers = c("6", "7", "171", "173", "175"), Operators_Assembly_Workers = c("8"), Unskilled_Workers = c("9", "192", "193", "194"), Armed_Forces = "10", Other_Unknown = c("90", "99")))

student_data‘$Father's occupation‘ <- factor(student_data$Father's occupation, levels = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 90, 99, 122, 123, 125, 131, 132, 134, 141, 143, 144, 151, 152, 153, 171, 173, 175, 191, 192, 193, 194))

Collapse the levels into broader categories

student_data <- mutate(student_data, Father's occupation = fct_collapse(Father's occupation, Student = c("0"), High_Level_Professionals = c("1", "2", "122", "123", "125"), Intermediate_Professionals = c("3", "131", "132", "134"), Administrative_Staff = c("4", "141", "143", "144"), Service_Workers = c("5", "151", "152", "153", "191"), Skilled_Workers = c("6", "7", "171", "173", "175"), Operators_Assembly_Workers = c("8"), Unskilled_Workers = c("9", "192", "193", "194"), Armed_Forces = "10", Other_Unknown = c("90", "99")))

student_data‘$MaritalStatus$‘ < $-factor(student_data$`Marital Status`, levels = c(“1”, “2”, “3”, “4”, “5”, “6”))

student_data‘$MaritalStatus$‘ < $-fct_recode(student_data$`Marital Status`, “Single” = “1”, “Married” = “2”, “Other” = “3”,
“Other” = “4”,
“Other” = “5”,
“Other” = “6”)

student_data <- student_data %>% mutate(`Curricular units all year (enrolled)` = (`Curricular units 1st sem (enrolled)` + `Curricular units 2nd sem (enrolled)`) / 2, `Curricular units all year (evaluations)` = (`Curricular units 1st sem (evaluations)` + `Curricular units 2nd sem (evaluations)`) / 2, `Curricular units all year (approved)` = (`Curricular units 1st sem (approved)` + `Curricular units 2nd sem (approved)`) / 2, `Curricular units all year (grade)` = (`Curricular units 1st sem (grade)` + `Curricular units 2nd sem (grade)`) / 2)

Remove semesterly data - Don’t run twice

student_data <- student_data[,-c(22,23,24,25,26,27,28,29,30,31,32,33)]

names(student_data) <- make.names(names(student_data)) student_data$Target <- factor(student_data$Target)

“‘

Introduction

ipsum loren

Methods

ipsum loren

Exploratory Data Analysis

ipsum loren

Including Plots

ipsum loren