Credit score approval prediction

Jan Alexander Jensen

2020/4/16

Handling nominal variable with dummy variable

Use library 'fastDummies' to handle the nominal variable $OCCUPATION_TYPE.

```
library(fastDummies)
library(janitor)

dummy_var <- function(data){

  # Since the library 'fastDummies' tansforms all factor variable into dummy variables,
  # we will convert our target "y" (factor) into a character variable
  # to avoid it being transformed to dummy variable.

  data$y <- as.numeric(as.character(data$y))

  # transform all factor variables to dummy variables,
  # and removes the original variables that were used to generate the dummy variables.
  data_dummy <- fastDummies::dummy_cols(data, remove_selected_columns=TRUE)

  # column name convention fix (mlr3 name convention - space to underscore)
  data_dummy <- clean_names(data_dummy)


  data_dummy <- as.data.frame(sapply(data_dummy, as.numeric))
  data_dummy$y <- as.factor(data_dummy$y)

  dummy_var <- data_dummy
}

# ------------------------------------------
# ----------------------- handle missing data
# ----------------------- OCCUPATION_TYPE
# ------------------------------------------

setwd("C:/Users/user/Documents/R-projects/i2ml_final_project")

dl_dummy_data <- read.csv2("credit_card_prediction/dl_na_data.csv", header = TRUE)
dl_dummy_data <- dummy_var(dl_dummy_data)

mf_dummy_data <- read.csv2("credit_card_prediction/mf_na_data.csv", header = TRUE)
mf_dummy_data <- dummy_var(mf_dummy_data)

mice_dummy_data <- read.csv2("credit_card_prediction/mice_na_data.csv", header = TRUE)
mice_dummy_data <- dummy_var(mice_dummy_data)

# row, column, NA
cat("dl_dummy_data\t", dim(dl_dummy_data), any(is.na(dl_dummy_data)), "\n")
cat("mf_dummy_data\t", dim(mf_dummy_data), any(is.na(mf_dummy_data)), "\n")
```

```r
cat("mice_dummy_data\t", dim(mice_dummy_data), any(is.na(mice_dummy_data)), "\n")

# store data with dummy variable
# write_csv2(dl_dummy_data,"credit_card_prediction/dummy_data/dl_dummy_data.csv")
# write_csv2(mf_dummy_data,"credit_card_prediction/dummy_data/mf_dummy_data.csv")
# write_csv2(mice_dummy_data,"credit_card_prediction/dummy_data/mice_dummy_data.csv")
```

```
## dl_dummy_data     25134 55 FALSE
## mf_dummy_data     36457 55 FALSE
## mice_dummy_data   36457 55 FALSE
```

Load all data for training (one-hot, dummy, IV)

```r
setwd("C:/Users/user/Documents/R-projects/i2ml_final_project")
library(mlr3)

# function to load data into task and define target
dataToTask <- function(path, id, sep=';', header=TRUE){
  dt <- read.csv2(path, sep = sep, header = header)
  dt <- as.data.frame(sapply(dt, as.numeric))
  dt$y <- as.factor(dt$y)
  dataToTask <- TaskClassif$new(id = id, backend = dt, target = "y")
}

dl_dummy_task <-
  dataToTask("credit_card_prediction/dummy_data/dl_dummy_data.csv", "dl_dummy")
dl_oh_task <-
  dataToTask("credit_card_prediction/oh_data/dl_oh_data.csv", "dl_oh", sep = ',')
dl_iv_task <-
  dataToTask("credit_card_prediction/iv_data/dl_iv_data.csv", "dl_iv")

mf_dummy_task <-
  dataToTask("credit_card_prediction/dummy_data/mf_dummy_data.csv", "mf_dummy")
mf_oh_task <-
  dataToTask("credit_card_prediction/oh_data/mf_oh_data.csv", "mf_oh", sep = ',')
mf_iv_task <-
  dataToTask("credit_card_prediction/iv_data/mf_iv_data.csv", "mf_iv")

mice_dummy_task <-
  dataToTask("credit_card_prediction/dummy_data/mice_dummy_data.csv", "mice_dummy")
mice_oh_task <-
  dataToTask("credit_card_prediction/oh_data/mice_oh_data.csv", "mice_oh", sep = ',')
mice_iv_task <-
  dataToTask("credit_card_prediction/iv_data/mice_iv_data.csv", "mice_iv")

# combine all tasks into one list
dl <- list(dummy=dl_dummy_task, oh=dl_oh_task, iv=dl_iv_task)
mf <- list(dummy=mf_dummy_task, oh=mf_oh_task, iv=mf_iv_task)
mice <- list(dummy=mice_dummy_task, oh=mice_oh_task, iv=mice_iv_task)


# tasks[["<type>"]][["<code>"]], tasks$<type>$<code>
# ex. tasks[["dl"]][["dummy"]], tasks$dl$dummy
```

```r
tasks <- list(dl=dl, mf=mf, mice=mice)

# remove unused variables (save memory)
rm(dl, mf, mice)
rm(dl_dummy_task, mf_dummy_task, mice_dummy_task)
rm(dl_oh_task, mf_oh_task, mice_oh_task)
rm(dl_iv_task, mf_iv_task, mice_iv_task)

# print task ids and data size
for(t in tasks){
  for(c in t){
    cat(c$id, dim(c$data()), "\n")
  }
}
```

```
## dl_dummy 25134 55
## dl_oh 25134 57
## dl_iv 25134 35
## mf_dummy 36457 55
## mf_oh 36457 57
## mf_iv 36457 35
## mice_dummy 36457 55
## mice_oh 36457 57
## mice_iv 36457 35
```

KNN