Lab block 2

johmo870, nisra674

2024-12-07

Statement of Contribution

- Nisal Amashan(nisra674) - Assignment 1, Assignment 3, Report
- John Möller (johmo870) - Assignment 2, Assignment 4, Report

Assignment 2

Data import and cleaning

The categories specified in the data compared to the instructions differed. Here's a summary of the differences:

- There are a number of names without dashes in the instructions but that have dashes in the csv, those are: blue-collar, self-employed.
- A number of columns had unknown as a category, specifically those between marital to to contact and poutcome.
- poutcome had 'other' in csv which is assumed in this report to be the same as 'nonexistent' given this category doesn't come up.
- Day of week column does not exist in csv.
- There exists a column called "day" which is not specified in the instruction. It is presumed in the report that this is the numbered day of the month.
- Not previously contacted in pdays is encoded as -1 in the csv as opposed to 999 as it is described in the instructions.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
df_raw <- read_delim(file = "data/bank-full.csv", delim = ";")
```

```
## Rows: 45211 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ";"
## chr (10): job, marital, education, default, housing, loan, contact, month, p...
## dbl  (7): age, balance, day, duration, campaign, pdays, previous
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df <- df_raw
df$job <- factor(df$job, levels = c('admin.','blue-collar','entrepreneur','housemaid','management','reti
df$marital <- factor(df$marital, levels = c('divorced','married','single','unknown'))
df$education <- factor(df$education, levels = c('primary', 'secondary', 'tertiary', 'unknown'))
df$default <- factor(df$default, levels = c('no', 'yes', 'unknown'))
df$housing <- factor(df$housing, levels = c('no', 'yes', 'unknown'))
df$loan <- factor(df$loan, levels = c('no', 'yes', 'unknown'))
df$contact <- factor(df$contact, levels = c('cellular', 'telephone', 'unknown'))
df$month <- factor(df$month, levels = c('jan','feb', 'mar', 'apr', 'may', 'jun', 'jul', 'aug', 'sep', '
#df$day <- factor(df$day, levels = c('mon', 'tue', 'wed', 'thu', 'fri', 'unknown'))
df$poutcome <- factor(df$poutcome, levels = c('failure', 'other', 'success', 'unknown'))
colSums(is.na(df))
```

```
##       age       job   marital education   default   balance   housing      loan
##         0         0         0         0         0         0         0         0
##   contact       day     month  duration  campaign     pdays  previous  poutcome
##         0         0         0         0         0         0         0         0
##         y
##         0
```

Task 1

Let's divide the data as specified in lecture 2a.

```
data <- df
n=dim(data)[1]
set.seed(12345)
id=sample(1:n, floor(n*0.4))
train=data[id,]
id1=setdiff(1:n, id)
set.seed(12345)
id2=sample(id1, floor(n*0.3))
valid=data[id2,]
id3=setdiff(id1,id2)
test=data[id3,]
```