hw3.R

joann

2021-02-20

```
#Read and inspect data set
data <- read.csv("C:/Users/joann/OneDrive/Desktop/missing data/week 2/aug_train.csv", na.strings = "")
str(data)
```

```
## 'data.frame':    19158 obs. of  14 variables:
##  $ enrollee_id           : int  8949 29725 11561 33241 666 21651 28806 402 27107 699 ...
##  $ city                  : chr  "city_103" "city_40" "city_21" "city_115" ...
##  $ city_development_index: num  0.92 0.776 0.624 0.789 0.767 0.764 0.92 0.762 0.92 0.92 ...
##  $ gender                : chr  "Male" "Male" NA NA ...
##  $ relevent_experience   : chr  "Has relevent experience" "No relevent experience" "No relevent expe
##  $ enrolled_university   : chr  "no_enrollment" "no_enrollment" "Full time course" NA ...
##  $ education_level       : chr  "Graduate" "Graduate" "Graduate" "Graduate" ...
##  $ major_discipline      : chr  "STEM" "STEM" "STEM" "Business Degree" ...
##  $ experience            : chr  ">20" "15" "5" "<1" ...
##  $ company_size          : chr  NA "50-99" NA NA ...
##  $ company_type          : chr  NA "Pvt Ltd" NA "Pvt Ltd" ...
##  $ last_new_job          : chr  "1" ">4" "never" "never" ...
##  $ training_hours        : int  36 47 83 52 8 24 24 18 46 123 ...
##  $ target                : num  1 0 0 1 0 1 0 1 1 0 ...
```

```
#check for missing values
sapply(data, function(x) sum(is.na(x)))
```

```
##            enrollee_id                   city city_development_index
##                      0                      0                      0
##                 gender    relevent_experience    enrolled_university
##                   4508                      0                    386
##        education_level       major_discipline             experience
##                    460                   2813                     65
##           company_size           company_type           last_new_job
##                   5938                   6140                    423
##         training_hours                 target
##                      0                      0
```

```
#the variables contain missing values are all categorical


#Encode character variables

unique(data$relevent_experience )
```

```
## [1] "Has relevent experience" "No relevent experience"
```

```
library(plyr)
```

```
## Warning: package ’plyr’ was built under R version 4.0.3
```

```
data$relevent_experience <- revalue(data$relevent_experience, c("Has relevent experience"=1))
data$relevent_experience <- revalue(data$relevent_experience, c("No relevent experience"=0))
data$relevent_experience <-as.numeric(data$relevent_experience)

unique(data$last_new_job)
```

```
## [1] "1"     ">4"    "never" "4"     "3"     "2"     NA
```

```
data$last_new_job <- revalue(data$last_new_job, c("never"=0))
data$last_new_job <- revalue(data$last_new_job, c(">4"=5))
data$last_new_job <-as.numeric(data$last_new_job)

unique(data$enrolled_university )
```

```
## [1] "no_enrollment"    "Full time course" NA                 "Part time course"
```

```
data$enrolled_university <- revalue(data$enrolled_university, c("no_enrollment"=0))
data$enrolled_university <- revalue(data$enrolled_university, c("Part time course"=1))
data$enrolled_university <- revalue(data$enrolled_university,c("Full time course" = 2))
data$enrolled_university <-as.numeric(data$enrolled_university)

unique(data$education_level)
```

```
## [1] "Graduate"       "Masters"        "High School"    NA
## [5] "Phd"            "Primary School"
```

```
data$education_level <- as.numeric(factor(data$education_level,
                                          levels = c("Primary School",
                                                     "High School","Graduate",
                                                     "Masters","Phd")))
unique(data$gender)
```

```
## [1] "Male"   NA       "Female" "Other"
```

```
data$gender <- as.factor(data$gender)

#I will keep the variables that can be used for my analysis
library(dplyr)
```

```
##
## Attaching package: ’dplyr’
## The following objects are masked from ’package:plyr’:
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
data2 = select(data,’city_development_index’,’training_hours’,’gender’,’relevent_experience’,
              ’last_new_job’,’enrolled_university’,’education_level’,’target’)

#Generate missing values for training_hours depending on one variable
library(dplyr)
data_new = select(data2,’city_development_index’,’training_hours’)
library(mice)
```

```
## Warning: package ’mice’ was built under R version 4.0.3

##
## Attaching package: ’mice’

## The following object is masked from ’package:stats’:
##
##     filter

## The following objects are masked from ’package:base’:
##
##     cbind, rbind
```

```
cont_cat = ampute(data_new,prop = 0.2,patterns=c(1,0),mech = "MAR")$amp
data2[’training_hours’] = cont_cat[’training_hours’]

#check again for the generated missing values
sapply(data2, function(x) sum(is.na(x)))
```

```
## city_development_index         training_hours                 gender
##                      0                   3801                   4508
##    relevent_experience           last_new_job    enrolled_university
##                      0                    423                    386
##        education_level                 target
##                    460                      0
```

```
# regression imputation for the numeric variable
#variables without missing values are: target and city_development_index
data3 = select(data2,’city_development_index’,’training_hours’,’target’)
data3 =complete(mice(data.frame(data3), method = "norm.predict", m = 1, maxit = 1))
```

```
##
##  iter imp variable
##   1   1  training_hours
```

```
data2[’training_hours’]= data3[’training_hours’]


#listwise deletion for the missing categorical values
data2 = na.omit(data2)

#original complete data set
data_complete = na.omit(data)

#Linear regression analysis for the target variable
#0-not looking for a job change 1-looking for a job change
#model with the resulting data set
model1 = lm(target ~ city_development_index+training_hours+gender+relevent_experience+
            last_new_job+enrolled_university+education_level, data = data2)
summary(model1)
```

```
##
## Call:
## lm(formula = target ~ city_development_index + training_hours +
##     gender + relevent_experience + last_new_job + enrolled_university +
##     education_level, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.77511 -0.22337 -0.11791 -0.08057  0.95391
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.202e+00  3.254e-02  36.944  < 2e-16 ***
## city_development_index -1.106e+00  2.920e-02 -37.889  < 2e-16 ***
## training_hours         -1.249e-04  6.205e-05  -2.013  0.04411 *
## genderMale             -3.366e-02  1.195e-02  -2.817  0.00485 **
## genderOther             3.372e-03  3.210e-02   0.105  0.91635
## relevent_experience    -9.390e-02  8.617e-03 -10.897  < 2e-16 ***
## last_new_job            1.928e-03  2.091e-03   0.922  0.35648
## enrolled_university     3.767e-02  4.827e-03   7.804  6.4e-15 ***
## education_level         1.536e-02  5.086e-03   3.019  0.00254 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.395 on 14152 degrees of freedom
## Multiple R-squared:  0.1207, Adjusted R-squared:  0.1202
## F-statistic: 242.7 on 8 and 14152 DF,  p-value: < 2.2e-16
```

```
#model with the original complete data cases
model2 = lm(target ~ city_development_index+training_hours+gender+relevent_experience+
              last_new_job+enrolled_university+education_level, data = data_complete)
summary(model2)
```

```
##
```

```
## Call:
## lm(formula = target ~ city_development_index + training_hours +
##     gender + relevent_experience + last_new_job + enrolled_university +
##     education_level, data = data_complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72917 -0.11579 -0.06287 -0.04783  0.99140
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.357e+00  3.823e-02  35.503  < 2e-16 ***
## city_development_index -1.340e+00  3.138e-02 -42.714  < 2e-16 ***
## training_hours         -9.011e-05  5.909e-05  -1.525 0.127303
## genderMale             -8.893e-03  1.249e-02  -0.712 0.476614
## genderOther             2.687e-02  4.000e-02   0.672 0.501729
## relevent_experience    -3.748e-02  1.109e-02  -3.381 0.000726 ***
## last_new_job            2.277e-04  2.183e-03   0.104 0.916921
## enrolled_university     1.613e-02  6.064e-03   2.660 0.007827 **
## education_level        -5.279e-03  6.807e-03  -0.776 0.438046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3366 on 8946 degrees of freedom
## Multiple R-squared:  0.1807, Adjusted R-squared:   0.18
## F-statistic: 246.7 on 8 and 8946 DF,  p-value: < 2.2e-16
```

```
#Comparing the two results:
#the variables training hours and education level are significant in the resulting
#dataset but not significant in the complete data set
#However, the R-square value is larger for the complete data set
#this implies that the linear regression model fits better for the complete data set
```