# HW 9

August 3, 2021

# 1 IST 387 HW 9



[1]:
```
# Enter your name here: Ezra Cohen
```

### 1.0.1 Attribution statement: (choose only one and delete the rest)

[2]:
```
# 1. I did this homework by myself, with help from the book and the professor.
```

**Association mining** can be applied to many data problems beyond the well-known example of **finding relationships between different products in customer shopping data**. In this homework assignment, we will explore **real data** from the banking sector and look for **patterns associated with the likelihood of responding positively to a direct marketing campaign and signing up for a term deposit with the bank (stored in the variable "y")**. You can find out more about the variables in this dataset here: https://archive.ics.uci.edu/ml/datasets/bank+marketing

## 1.1 Part 1: Explore Data Set

A. Copy the contents of the following URL to a dataframe called bank: https://ist387.s3.us-east-2.amazonaws.com/data/bank-full.csv

**Hint**: Even though this is a .csv file, chances are R won't be able to read it in correctly using the read_csv() function. If you take a closer look at the contents of the URL file, you may notice each field is separated by a **semicolon** (;) rather than a comma. In situations like this, consider using something like this:

[3]:
```
url<-"https://ist387.s3.us-east-2.amazonaws.com/data/bank-full.csv"
bank <- read.table(url, sep=";", header = TRUE)
dim(bank)
bank
```

1. 41188 2. 21

A data.frame: 41188 × 21

| age | job | marital | education | default | housing | loan |
|---|---|---|---|---|---|---|
| <int> | <chr> | <chr> | <chr> | <chr> | <chr> | <chr> |
| 56 | housemaid | married | basic.4y | no | no | no |
| 57 | services | married | high.school | unknown | no | no |
| 37 | services | married | high.school | no | yes | no |
| 40 | admin. | married | basic.6y | no | no | no |
| 56 | services | married | high.school | no | no | yes |
| 45 | services | married | basic.9y | unknown | no | no |
| 59 | admin. | married | professional.course | no | no | no |
| 41 | blue-collar | married | unknown | unknown | no | no |
| 24 | technician | single | professional.course | no | yes | no |
| 25 | services | single | high.school | no | yes | no |
| 41 | blue-collar | married | unknown | unknown | no | no |
| 25 | services | single | high.school | no | yes | no |
| 29 | blue-collar | single | high.school | no | no | yes |
| 57 | housemaid | divorced | basic.4y | no | yes | no |
| 35 | blue-collar | married | basic.6y | no | yes | no |
| 54 | retired | married | basic.9y | unknown | yes | yes |
| 35 | blue-collar | married | basic.6y | no | yes | no |
| 46 | blue-collar | married | basic.6y | unknown | yes | yes |
| 50 | blue-collar | married | basic.9y | no | yes | yes |
| 39 | management | single | basic.9y | unknown | no | no |
| 30 | unemployed | married | high.school | no | no | no |
| 55 | blue-collar | married | basic.4y | unknown | yes | no |
| 55 | retired | single | high.school | no | yes | no |
| 41 | technician | single | high.school | no | yes | no |
| 37 | admin. | married | high.school | no | yes | no |
| 35 | technician | married | university.degree | no | no | yes |
| 59 | technician | married | unknown | no | yes | no |
| 39 | self-employed | married | basic.9y | unknown | no | no |
| 54 | technician | single | university.degree | unknown | no | no |
| 55 | unknown | married | university.degree | unknown | unknown | unknow |
| 35 | technician | divorced | basic.4y | no | no | no |
| 35 | technician | divorced | basic.4y | no | yes | no |
| 33 | admin. | married | university.degree | no | no | no |
| 33 | admin. | married | university.degree | no | yes | no |
| 60 | blue-collar | married | basic.4y | no | yes | no |
| 35 | technician | divorced | basic.4y | no | yes | no |
| 54 | admin. | married | professional.course | no | no | no |
| 38 | housemaid | divorced | university.degree | no | no | no |
| 32 | admin. | married | university.degree | no | no | no |
| 32 | admin. | married | university.degree | no | yes | no |
| 38 | entrepreneur | married | university.degree | no | no | no |
| 62 | services | married | high.school | no | yes | no |
| 40 | management | divorced | university.degree | no | yes | no |
| 33 | student | married | professional.course | no | yes | no |
| 31 | admin. | single | university.degree | no | yes | no |
| 62 | retired | married | university.degree | no | yes | no |
| 62 | retired | married | university.degree | no | yes | no |
| 34 | student | single | unknown | no | yes | no |
| 38 | housemaid | divorced | high.school | no | yes | yes |
| 57 | retired | married | professional.course | no | yes | no |

Make sure there are **41,188** rows and **21** columns in your **bank** df.

B. Next, we will focus on some key factor variables from the dataset, and convert a few numeric ones to factor variables. Execute the following commands and write a comment describing how the conversion for each numeric variable works and what the variables in the resulting dataframe are.

[4]:
```
bank_new <- data.frame(job=bank$job,
                       marital=bank$marital,
                       housing_loan=bank$housing,
                       young=as.factor((bank$age<median(bank$age))),
#Makes it into a factor based on if the person is older or younger than the␣
↪median of bank$age, if they are younger it is true if they are older it is␣
↪false
                       contacted_more_than_once=as.factor((bank$campaign>1)),
#Makes it into a factor based on if they were contacted more than one times if␣
↪they were contacted 1 or last times then it would be false and if they were␣
↪contacted more than once it would be true
                       contacted_before_this_campaign=as.
↪factor((bank$previous<0)),
#Makes it into a factor based on if they had less than zero Banks prior to␣
↪this, and I don't understand the point of this because that doesn't seem␣
↪possible and the entire column is just false
                       success=(bank$y))
bank_new
#Job is what job they do, marital is if they are married or not, housing loan␣
↪is whether they have a loan on their house, young is if they are young or␣
↪not, contacted more than once is if the bank contacted them more than once,␣
↪Contacted before this campaign is if they had less than zero banks prior to␣
↪this, and success is number of successful term deposit sign-up
```

A data.frame: 41188 × 7

| job | marital | housing_loan | young | contacted_more_than_once | con |
|---|---|---|---|---|---|
| <chr> | <chr> | <chr> | <fct> | <fct> | <fc |
| housemaid | married | no | FALSE | FALSE | FAI |
| services | married | no | FALSE | FALSE | FAI |
| services | married | yes | TRUE | FALSE | FAI |
| admin. | married | no | FALSE | FALSE | FAI |
| services | married | no | FALSE | FALSE | FAI |
| services | married | no | FALSE | FALSE | FAI |
| admin. | married | no | FALSE | FALSE | FAI |
| blue-collar | married | no | FALSE | FALSE | FAI |
| technician | single | yes | TRUE | FALSE | FAI |
| services | single | yes | TRUE | FALSE | FAI |
| blue-collar | married | no | FALSE | FALSE | FAI |
| services | single | yes | TRUE | FALSE | FAI |
| blue-collar | single | no | TRUE | FALSE | FAI |
| housemaid | divorced | yes | FALSE | FALSE | FAI |
| blue-collar | married | yes | TRUE | FALSE | FAI |
| retired | married | yes | FALSE | FALSE | FAI |
| blue-collar | married | yes | TRUE | FALSE | FAI |
| blue-collar | married | yes | FALSE | FALSE | FAI |
| blue-collar | married | yes | FALSE | FALSE | FAI |
| management | single | no | FALSE | FALSE | FAI |
| unemployed | married | no | TRUE | FALSE | FAI |
| blue-collar | married | yes | FALSE | FALSE | FAI |
| retired | single | yes | FALSE | FALSE | FAI |
| technician | single | yes | FALSE | FALSE | FAI |
| admin. | married | yes | TRUE | FALSE | FAI |
| technician | married | no | TRUE | FALSE | FAI |
| technician | married | yes | FALSE | FALSE | FAI |
| self-employed | married | no | FALSE | FALSE | FAI |
| technician | single | no | FALSE | TRUE | FAI |
| unknown | married | unknown | FALSE | FALSE | FAI |
| | | | | | |
| technician | divorced | no | TRUE | FALSE | FAI |
| technician | divorced | yes | TRUE | FALSE | FAI |
| admin. | married | no | TRUE | FALSE | FAI |
| admin. | married | yes | TRUE | FALSE | FAI |
| blue-collar | married | yes | FALSE | TRUE | FAI |
| technician | divorced | yes | TRUE | TRUE | FAI |
| admin. | married | no | FALSE | TRUE | FAI |
| housemaid | divorced | no | FALSE | TRUE | FAI |
| admin. | married | no | TRUE | FALSE | FAI |
| admin. | married | yes | TRUE | TRUE | FAI |
| entrepreneur | married | no | FALSE | TRUE | FAI |
| services | married | yes | FALSE | TRUE | FAI |
| management | divorced | yes | FALSE | TRUE | FAI |
| student | married | yes | TRUE | FALSE | FAI |
| admin. | single | yes | TRUE | FALSE | FAI |
| retired | married | yes | FALSE | FALSE | FAI |
| retired | married | yes | FALSE | FALSE | FAI |
| student | single | yes | TRUE | FALSE | FAI |
| housemaid | divorced | yes | FALSE | FALSE | FAI |
| retired | married | yes | FALSE | TRUE | FAI |

C. Count the number of successful term deposit sign-ups, using the table( ) command on the **success** variable.

```
[5]: table(bank_new$success)
```

```
   no   yes 
36548  4640 
```

D. Express the results of problem C as percentages by sending the results of the table( ) command into the prop.table( ) command.

```
[6]: prop.table(table(bank_new$success))
```

```
       no       yes 
0.8873458 0.1126542 
```

E. Using the same techniques, show the percentages for the **marital** and **housing_loan** variables as well.

```
[7]: prop.table(table(bank_new$marital))
prop.table(table(bank_new$housing_loan))
```

```
   divorced     married      single     unknown 
0.111974361 0.605224823 0.280858502 0.001942313 
```

```
        no    unknown        yes 
0.45212198 0.02403613 0.52384190 
```

## 1.2 Part 2: Coerce the data frame into transactions

F. Install and library two packages: **arules** and **arulesViz**.

```
[21]: #install.packages("arules")
#install.packages("arulesViz")
library(arules)
library(arulesViz)
```

G. Coerce the **bank_new** dataframe into a **sparse transactions matrix** called **bankX**.

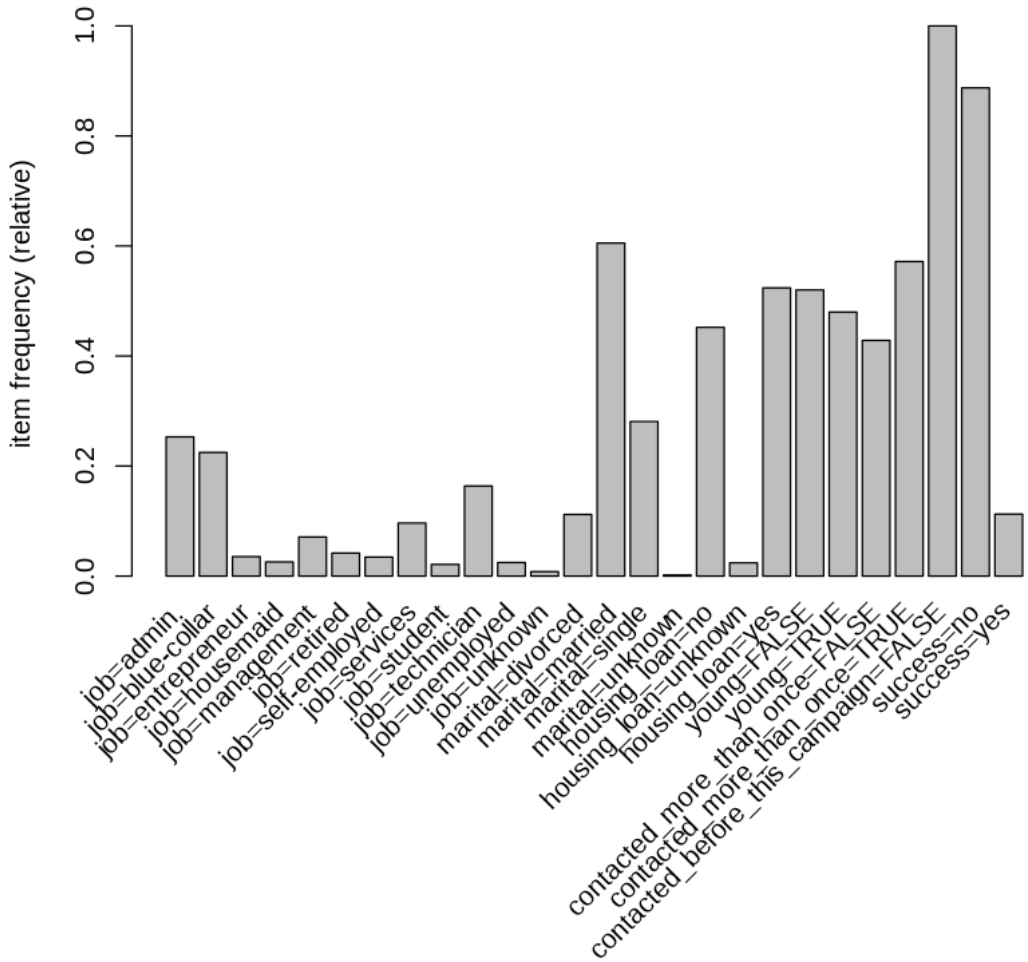
```
[9]: bankx<-as(bank_new,"transactions")
```

```
Warning message:
"Column(s) 1, 2, 3, 7 not logical or factor. Applying default discretization
(see '? discretizeDF')."
```

H. Use the itemFrequency( ) and itemFrequencyPlot( ) commands to explore the contents of **bankX**. What do you see?

[10]:
```
itemFrequency(bankx)
itemFrequencyPlot(bankx)
#From this we can see what occurred in what percent of the cases and we can get
↪information like a majority of the people have a job as admin, most of them
↪have a marital status of married, and and way more people have success as a
↪no then as a yes
```

**job=admin.** 0.253034864523648 **job=blue-collar** 0.224677090414684 **job=entrepreneur** 0.035350101971448 **job=housemaid** 0.0257356511605322 **job=management** 0.0709915509371662 **job=retired** 0.0417597358453919 **job=self-employed** 0.0345003399048266 **job=services** 0.0963630183548606 **job=student** 0.0212440516655336 **job=technician** 0.16371273186365 **job=unemployed** 0.0246188210158299 **job=unknown** 0.00801204234242983 **marital=divorced** 0.111974361464504 **marital=married** 0.605224822763912 **marital=single** 0.280858502476449 **marital=unknown** 0.00194231329513451 **housing_loan=no** 0.452121977274934 **housing_loan=unknown** 0.0240361270272895 **housing_loan=yes** 0.523841895697776 **young=FALSE** 0.520054384772264 **young=TRUE** 0.479945615227736 **contacted_more_than_once=FALSE** 0.428328639409537 **contacted_more_than_once=TRUE** 0.571671360590463 **contacted_before_this_campaign=FALSE** 1 **success=no** 0.887345828882199 **success=yes** 0.112654171117801




I. This is a fairly large dataset, so we will explore only the first 10 observations in the **bankX** transaction matrix:

[11]:
```
inspect(bankx[1:10])
```

```
     items                                       transactionID
[1]  {job=housemaid,
      marital=married,
      housing_loan=no,
      young=FALSE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                            1
[2]  {job=services,
```

```
      marital=married,
      housing_loan=no,
      young=FALSE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                  2
[3]  {job=services,
      marital=married,
      housing_loan=yes,
      young=TRUE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                  3
[4]  {job=admin.,
      marital=married,
      housing_loan=no,
      young=FALSE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                  4
[5]  {job=services,
      marital=married,
      housing_loan=no,
      young=FALSE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                  5
[6]  {job=services,
      marital=married,
      housing_loan=no,
      young=FALSE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                  6
[7]  {job=admin.,
      marital=married,
      housing_loan=no,
      young=FALSE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                  7
[8]  {job=blue-collar,
      marital=married,
      housing_loan=no,
      young=FALSE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                  8
```

```
[9]  {job=technician,
      marital=single,
      housing_loan=yes,
      young=TRUE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                      9
[10] {job=services,
      marital=single,
      housing_loan=yes,
      young=TRUE,
      contacted_more_than_once=FALSE,
      contacted_before_this_campaign=FALSE,
      success=no}                                      10
```

Explain the difference between **bank_new** and **bankX** in a block comment:

Bank_new is a data set and Bankx is a sparse transactions Matrix

### 1.3 Part 3: Use arules to discover patterns

**Support** is the proportion of times that a particular set of items occurs relative to the whole dataset. **Confidence** is proportion of times that the consequent occurs when the antecedent is present.

J. Use **apriori** to generate a set of rules with support over 0.005 and confidence over 0.3, and trying to predict who successfully signed up for a term deposit. **Hint:** You need to define the **right-hand side rule (rhs)**.

[50]:
```
ruleset<-apriori(bankx,
                 parameter=list(supp=0.006, conf=0.32),
                 control=list(verbose=F),
                 appearance=list(default="lhs",rhs=("success=yes")))
```

K. Use inspect() to review of the **ruleset**.

[51]:
```
inspect(ruleset)
```

```
    lhs                                       rhs               support
confidence   coverage     lift  count
[1] {job=student,
     marital=single}                       => {success=yes} 0.006409634
0.3203883 0.02000583 2.843999   264
[2] {job=student,
     marital=single,
     young=TRUE}                           => {success=yes} 0.006312518
0.3233831 0.01952025 2.870582   260
[3] {job=student,
     marital=single,
     contacted_before_this_campaign=FALSE} => {success=yes} 0.006409634
```

```
0.3203883 0.02000583 2.843999   264
[4] {job=student,
     marital=single,
     young=TRUE,
     contacted_before_this_campaign=FALSE} => {success=yes} 0.006312518
0.3233831 0.01952025 2.870582   260
```

L. Use the output of inspect( ) or inspectDT( ) and describe **any 2 rules** the algorithm found.

All of these rules are very similar so it would be difficult to discuss two separate rules so I will just discuss the last rule which includes everything from all the prior rules, this rule includes the job is student, the marital status is single, young is true, and contacted prior to Campaign is false, all of these things make sense together, someone who is a student is likely to be young and single and everyone was not contacted prior to the campaign so it would be false for everyone, and what all of this is implying is that younger less experienced people are more likely to sign up for it