# Assignment3

John Bute

2024-09-21

## Question 3

Before we begin, we must develop our code. My logic was that instead of computing the Confidence, Interest, … of every rule, why not create a function (calculate_rule_metrics) that would a return a list of the calculated metrics of rule A->B. Then, I created a small database of the rules to study, looped through them, and appended the results to a dataframe as seen below. From there, we can start evaluating each rule.

```
Transactions <- read.csv("C:/Users/johnb/Desktop/Machine Learning/data/Transactions.csv")

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

```
calculate_rule_metrics <- function(data, antecedent_cols, consequent_col) {
  if (!all(c(antecedent_cols, consequent_col) %in% names(data))) {
    stop("One or more specified columns do not exist in the dataframe.")
  }

  N <- nrow(data)

  freq_A <- data %>%
    filter(Reduce(`&`, lapply(antecedent_cols, function(col) data[[col]] == 1))) %>%
    nrow()

  freq_B <- sum(data[[consequent_col]] == 1)

  freq_A_and_B <- data %>%
    filter(Reduce(`&`, lapply(antecedent_cols, function(col) data[[col]] == 1)) & data[[consequent_col]]
    nrow()

  support <- freq_A_and_B / N
  confidence <- freq_A_and_B / freq_A
  interest <- confidence - (freq_B / N)
  lift <- (N^2 * support) / (freq_A * freq_B)
  conviction <- (1 - (freq_B / N)) / (1 - confidence)

  return(list(
    Support = support,
    Confidence = confidence,
    Interest = interest,
    Lift = lift,
    Conviction = conviction
  ))
}

results_df <- data.frame(
  Rule = character(),
  Support = numeric(),
  Confidence = numeric(),
  Interest = numeric(),
  Conviction = numeric(),
  stringsAsFactors = FALSE
)

rules <- list(
  list(antecedent = c("Red", "White"), consequent = "Green"),
  list(antecedent = c("Green"), consequent = "White"),
  list(antecedent = c("Red", "Green"), consequent = "White"),
  list(antecedent = c("Green"), consequent = "Red"),
  list(antecedent = c("Orange"), consequent = "Red"),
  list(antecedent = c("White", "Black"), consequent = "Yellow"),
  list(antecedent = c("Black"), consequent = "Green")
)

for (rule in rules){
  antecedent <- rule$antecedent
  consequent <- rule$consequent
  metrics <- calculate_rule_metrics(Transactions, antecedent, consequent)

  if(!is.null(metrics)){
    results_df <- rbind(
      results_df,
      data.frame(
        Rule = paste("{", paste(antecedent, collapse = ", "), "} -> {", consequent, "}"),
        Support = metrics$Support,
        Confidence = metrics$Confidence,
        Interest = metrics$Interest,
        Lift = metrics$Lift,
        Conviction = metrics$Conviction
      )
    )
  }
}
print(results_df)
```

```
##                              Rule    Support Confidence     Interest       Lift
## 1    { Red, White } -> { Green } 0.01120448 0.38095238  0.044817927  1.1333333
## 2         { Green } -> { White } 0.01540616 0.04583333 -0.012990196  0.7791667
## 3    { Red, Green } -> { White } 0.01120448 0.25806452  0.199240987  4.3870968
## 4           { Green } -> { Red } 0.04341737 0.12916667  0.011519608  1.0979167
## 5          { Orange } -> { Red } 0.00280112 0.03389831 -0.083748754  0.2881356
## 6 { White, Black } -> { Yellow } 0.03921569 0.90322581  0.852805638 17.9139785
## 7         { Black } -> { Green } 0.28851541 0.33279483 -0.003339623  0.9900646
##   Conviction
## 1  1.0723982
## 2  0.9863858
## 3  1.2685422
## 4  1.0132283
## 5  0.9133127
## 6  9.8123249
## 7  0.9949946
```

Among the 7 rules evaluated we have the following observations:

- Highest Lift: White and Black -> Yellow
  - Transactions with White and Black, that lead to yellow are positively correlated.
- Highest Confidence: White and Black -> Yellow
  - If a customer were to buy white and black, then we can strongly assume that they will also buy Yellow
- Highest Interest: White and Black -> Yellow
  - We can strongly assume that people who buy White and Black do not just randomly buy yellow too. Rather, it is a strong pattern.
- Highest Conviction: White and Black -> Yellow
  - A high conviction states that this rule is stronger than expected by random chance.

We will now proceed to inventing randomly 7 new rules:

Black -> Red | Orange, White, Red -> Green | Green -> Black | White -> Black | Red, Orange -> Yellow | Green, Black -> White | Red -> Black

Which we will now include in our analysis

```r
library(dplyr)
library(knitr)

calculate_rule_metrics <- function(data, antecedent_cols, consequent_col) {
  if (!all(c(antecedent_cols, consequent_col) %in% names(data))) {
    stop("One or more specified columns do not exist in the dataframe.")
  }

  N <- nrow(data)

  freq_A <- data %>%
    filter(Reduce(`&`, lapply(antecedent_cols, function(col) data[[col]] == 1))) %>%
    nrow()
  freq_B <- sum(data[[consequent_col]] == 1)

  # Calculate frequency of A and B occurring together
  freq_A_and_B <- data %>%
    filter(Reduce(`&`, lapply(antecedent_cols, function(col) data[[col]] == 1)) & data[[consequent_col]]
    nrow()

  support <- freq_A_and_B / N
  confidence <- freq_A_and_B / freq_A
  interest <- confidence - (freq_B / N)
  lift <- (N^2 * support) / (freq_A * freq_B)
  conviction <- (1 - (freq_B / N)) / (1 - confidence)

  return(list(
    Support = support,
    Confidence = confidence,
    Interest = interest,
    Lift = lift,
    Conviction = conviction
  ))
}

results_df <- data.frame(
  Rule = character(),
  Support = numeric(),
  Confidence = numeric(),
  Interest = numeric(),
  Conviction = numeric(),
  stringsAsFactors = FALSE
)

rules <- list(
  list(antecedent = c("Red", "White"), consequent = "Green"),
  list(antecedent = c("Green"), consequent = "White"),
  list(antecedent = c("Red", "Green"), consequent = "White"),
  list(antecedent = c("Green"), consequent = "Red"),
  list(antecedent = c("Orange"), consequent = "Red"),
  list(antecedent = c("White", "Black"), consequent = "Yellow"),
  list(antecedent = c("Black"), consequent = "Green"),
  list(antecedent = c("Black"), consequent = "Red"),
  list(antecedent = c("Orange", "White", "Red"), consequent = "Green"),
  list(antecedent = c("Green"), consequent = "Black"),
  list(antecedent = c("White"), consequent = "Black"),
  list(antecedent = c("Red", "Orange"), consequent = "Yellow"),
  list(antecedent = c("Green", "Black"), consequent = "White"),
  list(antecedent = c("Red"), consequent = "Black")

)

for (rule in rules){
  antecedent <- rule$antecedent
  consequent <- rule$consequent
  metrics <- calculate_rule_metrics(Transactions, antecedent, consequent)
  if(!is.null(metrics)){
    results_df <- rbind(
      results_df,
      data.frame(
        Rule = paste("{", paste(antecedent, collapse = ", "), "} -> {", consequent, "}"),
        Support = metrics$Support,
        Confidence = metrics$Confidence,
        Interest = metrics$Interest,
        Lift = metrics$Lift,
        Conviction = metrics$Conviction
      )
    )
  }
}

print(results_df)
```

```
##                                    Rule     Support Confidence     Interest
## 1          { Red, White } -> { Green } 0.011204482 0.38095238  0.044817927
## 2               { Green } -> { White } 0.015406162 0.04583333 -0.012990196
## 3          { Red, Green } -> { White } 0.011204482 0.25806452  0.199240987
## 4                 { Green } -> { Red } 0.043417367 0.12916667  0.011519608
## 5                { Orange } -> { Red } 0.002801120 0.03389831 -0.083748754
## 6       { White, Black } -> { Yellow } 0.039215686 0.90322581  0.852805638
## 7               { Black } -> { Green } 0.288515406 0.33279483 -0.003339623
## 8                 { Black } -> { Red } 0.092436975 0.10662359 -0.011023472
## 9  { Orange, White, Red } -> { Green } 0.001400560 1.00000000  0.663865546
## 10              { Green } -> { Black } 0.288515406 0.85833333 -0.008613445
## 11              { White } -> { Black } 0.043417367 0.73809524 -0.128851541
## 12       { Red, Orange } -> { Yellow } 0.001400560 0.50000000  0.449579832
## 13       { Green, Black } -> { White } 0.009803922 0.03398058 -0.024842947
## 14                { Red } -> { Black } 0.092436975 0.78571429 -0.081232493
##          Lift Conviction
## 1   1.1333333  1.0723982
## 2   0.7791667  0.9863858
## 3   4.3870968  1.2685422
## 4   1.0979167  1.0132283
## 5   0.2881356  0.9133127
## 6  17.9139785  9.8123249
## 7   0.9900646  0.9949946
## 8   0.9063005  0.9876609
## 9   2.9750000        Inf
## 10  0.9900646  0.9391992
## 11  0.8513732  0.5080214
## 12  9.9166667  1.8991597
## 13  0.5776699  0.9742832
## 14  0.9063005  0.6209150
```

Out of these rules, there are 4 that seem extremly useful to managers.

Firstly , { Green } -> { Black } is useful as its support and confidence says that this pattern occurs very frequently.

Secondly, { Black } -> { Green } is useful as its support and confidence says that this pattern occurs very frequently.

Thirdly, { White, Black } -> { Yellow } is extremly useful as its high confidence, lift, and conviction mean that this is a highly reliable pattern, where customers are likely to buy yellow items when buying white and black ones. Furthermore, the lift suggests this is a strong positive association.

Fourthly, { Red, Green } -> { White } is useful due to its lift, as you can use the rule to initiate bundling strategies.

NOTE: { Orange, White, Red } -> { Green } has very strong metrics, as its infinite conviction signifies that the action of buying green is never absent when buying orange, white and red. However, it might not be useful as this rule barely happens (0.0014 support value) and the categories are almost the entire domain (4 out of the 6 colors).

To determine a threshold for support, coverage, interest, and lift of rules derived from a given dataset, its important to know our dataset. For support, if it was a big dataset, any threshold that is low (0.01) is sufficient to take a look. However, in our case a 5% would be better to avoid uninteresting rules. For confidence, the situation is important, as for example, if you were using these rules to set up targeted promotions, you would want a higher confidence. For lift, you should look at all the rules you have discovered and determine what is a good cutoff to filter out rules that do not seem impactful.

Finally, interest, anything above 0 ensures that rules are not coincidental. Generally, you must first explore the situation, and then the data, to make sure you fully comprehend what rule and its metrics would be useful to you or the client in the situation. Finally, benchmarking would be a good idea, as comparing the rules with known patterns can ensure thresholds are not overly lenient or restrictive.