HW 3: Association Rules

Syracuse University: IST 707

Gavin Grosswald

Introduction

In the United States, investing and retirement planning have become integral aspects of personal finance, with a notable percentage of the population actively engaging in these activities. Access to investment opportunities has never been greater due to the increase in zero fee trading platforms such as Robinhood and Charles Schwab. However, despite the increased accessibility of investment opportunities, disparities in knowledge and participation persist, raising discussions on financial literacy and inclusivity within the United States.

Financial institutions and advisors aim to understand the drivers behind individual investment behaviors, particularly in relation to retirement planning and the adoption of financial products like the Personal Equity Plan (PEP). Through analyzing the demographic profiles and banking behaviors of individuals, the banks and advisors hope to uncover patterns that highlight the motivations and preferences guiding individuals' investment decisions.

The landscape of retirement planning and individual investing reflects broader societal shifts, highlighting the need for financial institutions to adapt and tailor their offerings to meet the evolving needs of diverse consumer groups. Through this exploration, banks and advisors aim to develop strategic approaches aimed at fostering financial inclusion and empowering individuals to navigate the complexities of personal finance with confidence and clarity.

Analysis, Models, and Results

About the Data

The Bank data was sourced from DePaul Univarity and contains 600 observations across 12 variables. The marketing team at a financial institution keeps records of customer data that can be leveraged when researching and launching new financial products. At present, the team is investigating a new Personal Equity Plan. To evaluate the success of the new product, the institution is mailing each customer a letter advertising the new product and tracking whether the customer responded and bought the new PEP product. Below is a table highlighting the provided information:

Variable	Description
ID	Unique identification number
Age	Customer age in years
Sex	Male / female
Region	Inner city / rural / suburban / town
Income	Customer income
Married	Customer marital status (yes / no)
Children	Number of children
Car	Customer car ownership status (yes / no)
Savings Account	Customer savings account status (yes / no)
Current Account	Customer current account status (yes / no)
Mortgage	Customer mortgage status (yes / no)
Personal Equity Plan	Customer personal equity plan (yes / no)

Table 1

Reading the Data

The Bank data csv file was read into R Studio and inserted into the “bank” dataframe. As part of the read.csv() function, all blank values were replaced with “NA”.

```
bank <- read.csv("C:\\Users\\gsgro\\OneDrive\\Desktop\\Syr_MSBA\\Term 3\\M
achine Learning\\bankdata_csv_all.csv", na.string = c(""))
```

The structure of the raw data was reviewed and is provided below:

```
## 'data.frame':    600 obs. of  12 variables:
##  $ id         : chr  "ID12101" "ID12102" "ID12103" "ID12104" ...
##  $ age        : int  48 40 51 23 57 57 22 58 37 54 ...
##  $ sex        : chr  "FEMALE" "MALE" "FEMALE" "FEMALE" ...
##  $ region     : chr  "INNER_CITY" "TOWN" "INNER_CITY" "TOWN" ...
##  $ income     : num  17546 30085 16575 20375 50576 ...
##  $ married    : chr  "NO" "YES" "YES" "YES" ...
##  $ children   : int  1 3 0 3 0 2 0 0 2 2 ...
##  $ car        : chr  "NO" "YES" "YES" "NO" ...
##  $ save_act   : chr  "NO" "NO" "YES" "NO" ...
##  $ current_act: chr  "NO" "YES" "YES" "YES" ...
##  $ mortgage   : chr  "NO" "YES" "NO" "NO" ...
##  $ pep        : chr  "YES" "NO" "NO" "NO" ...
```

Cleaning the Data

To prepare the data for associate rule mining, multiple transformations were required to convert the record data into transaction data:

Transformation 1: All numeric and character variables were converted to nominal as AR mining can only analyze nominal data.

Transformation 2: The children variable was converted to an ordinal factor.

Transformation 3: Duplicate values such as “YES” / “NO” were converted into unique values by concatenating the values with the variables name.

Transformation 4: The age and income variables were reviewed, and it was determined that due to their relatively even distributions, discretization would be beneficial for the overall analysis. The following bins were used for age and income:

```
Age: (0,20,30,40,50,60,100) → ("teens", "twenties", "thirties", "forties",
"fifties", "sixties")


##    teens twenties thirties  forties  fifties  sixties
##       21      123      117      141      100       98

Income: (0,15000,25000,35000,45000,100000) → ("0-14999", "15,000-24,999",
"25,000-34,999", "35,000-44999", "45,000+")


##       0-14999 15,000-24,999 25,000-34,999  35,000-44999       45,000+
##           102           200           142            82            74
```

The structure was then queried again and provided the following cleansed output:

```
## 'data.frame':    600 obs. of  12 variables:
##  $ id         : chr  "ID12101" "ID12102" "ID12103" "ID12104" ...
##  $ age        : Factor w/ 6 levels "teens","twenties",..: 4 4 5 2 5 5 2 5 3
5 ...
##  $ sex        : Factor w/ 2 levels "FEMALE","MALE": 1 2 1 1 1 1 2 2 1 2 ...
##  $ region     : Factor w/ 4 levels "INNER_CITY","RURAL",..: 1 4 1 4 2 4 2 4
3 4 ...
##  $ income     : Factor w/ 5 levels "0-14999","15,000-24,999",..: 2 3 2 2 5 4
1 2 3 2 ...
##  $ married    : Factor w/ 2 levels "married=NO","married=YES": 1 2 2 2 2 2 1
2 2 2 ...
##  $ children   : Ord.factor w/ 4 levels "0"<"1"<"2"<"3": 2 4 1 4 1 3 1 1 3 3
...
##  $ car        : Factor w/ 2 levels "car=NO","car=YES": 1 2 2 1 1 1 1 2 2 2 .
..
```

```
##  $ save_act   : Factor w/ 2 levels "save_act=NO",..: 1 1 2 1 2 2 1 2 1 2 ...
##  $ current_act: Factor w/ 2 levels "current_act=NO",..: 1 2 2 2 1 2 2 2 1 2
...
##  $ mortgage   : Factor w/ 2 levels "mortgage=NO",..: 1 2 1 1 1 1 1 1 1 1 ...
##  $ pep        : Factor w/ 2 levels "pep=NO","pep=YES": 2 1 1 1 1 2 2 1 1 1 .
..
```

Data Analysis

Each variable was plotted as seen in the bar charts below. From the analysis the following observations were made:

- Customers ages are evenly distributed except for those in their teens (ages 0-20)
- The banks clientele is made up of an even number of males and females, most of whom are married and have no children
- Most clients have income less than $35k
- Most clients have a savings account, and that account is managed by the bank
- Less than half of the bank's clientele have opted into the Personal Equity Plan

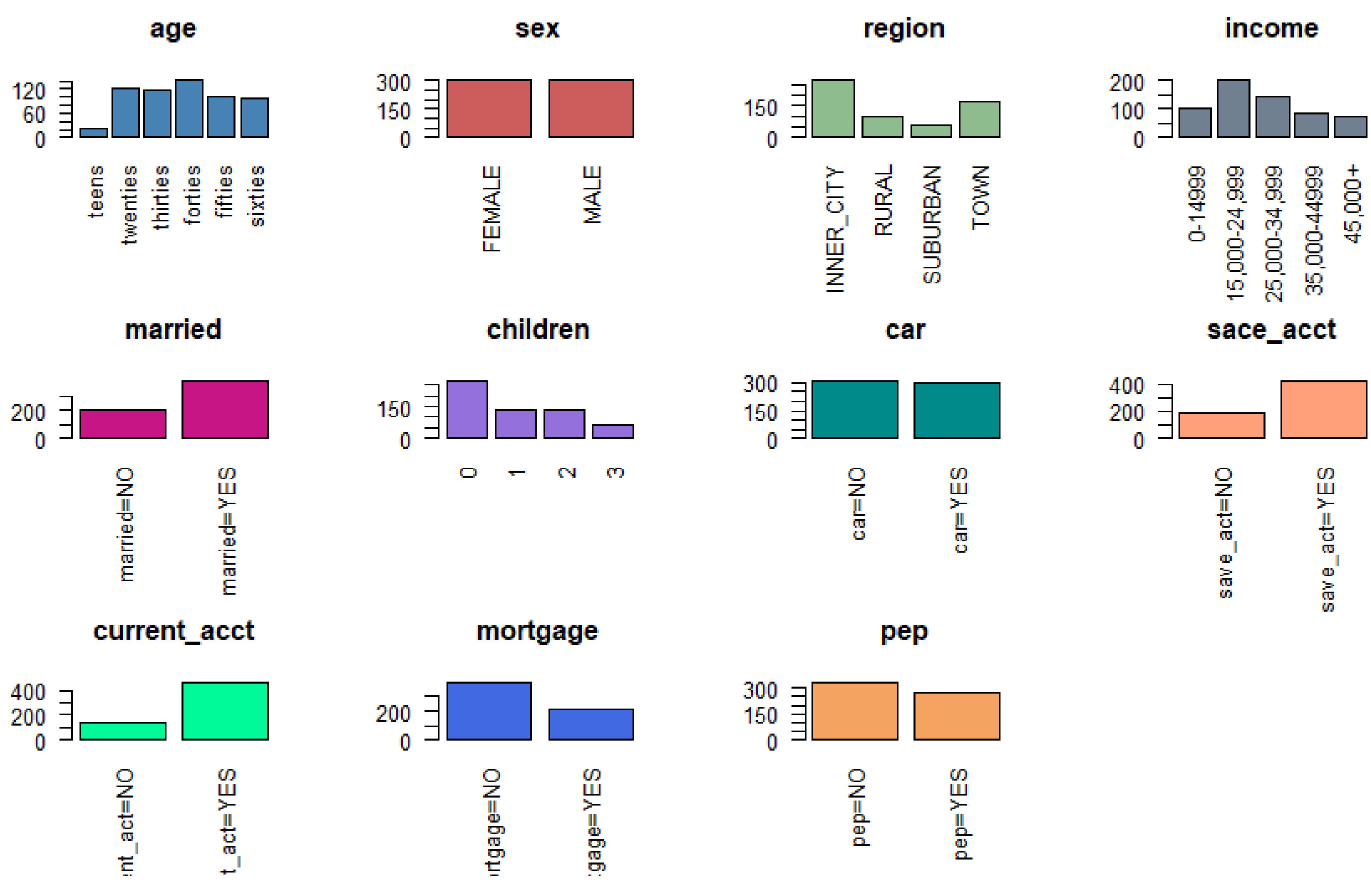

Table 2

Association Mining (Apriori)

The Apriori algorithm is helpful in data mining and association rule learning as it allows data scientists to uncover patterns within a dataset. It supports in identifying association rules through revealing common linkages (if this, then that) based on transactional data. Apriori can highlight frequent item sets and metrics such as support, confidence, and lift, which helps users to uncover significant associations. By leveraging Apriori and its outputs, users can gain insights into customer behavior and ultimately improve decision-making strategies.

Using the Apriori algorithm, the below frequency plot was generated which shows which sets of items are likely to appear together within a dataset.

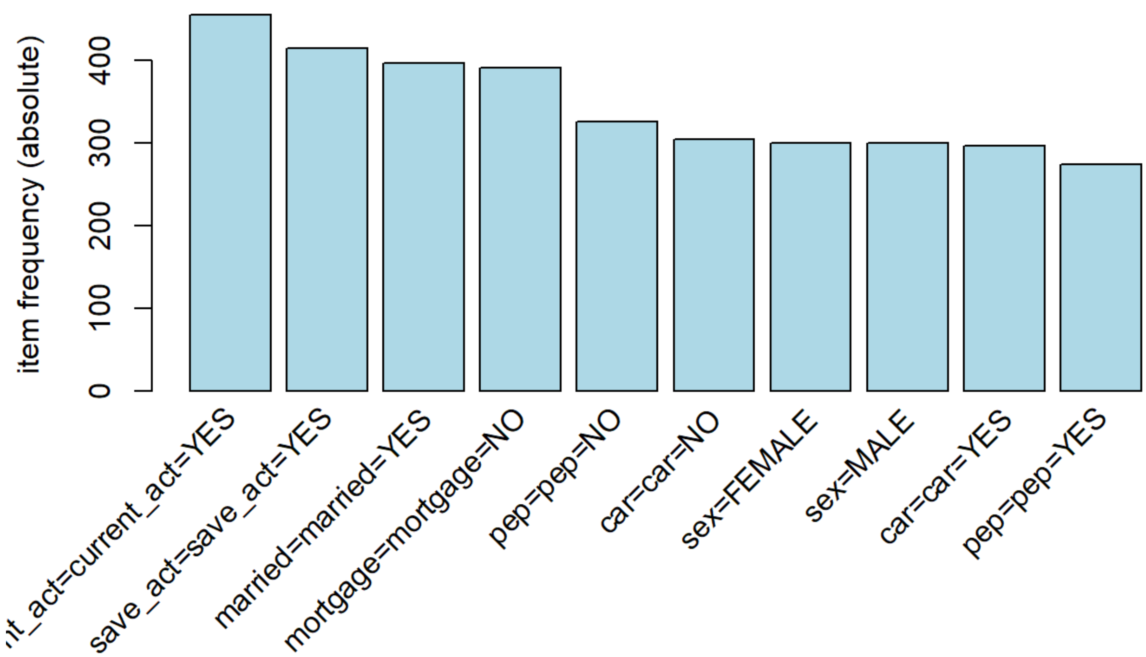

Table 3

The initial test to mine association rules utilized a support of 0.02 and a confidence of 0.9. The strongest result indicated that those in their 60s, who are female, own a car, have a savings account with the bank, and are enrolled in the PEP are likely to have an income of $45k plus. The lift was 8.11 indicating a strong relationship.

The goal, however, was to look at the PEP and what association mining rules were applicable to the PEP. Below are interesting rules where the PEP is set to be the value on the right-hand side:

1. {age=sixties, income=45,000+, children=1} → {pep=YES}
    a. Support = 0.02
    b. Confidence = 1.0
    c. Lift = 2.19

2. {sex=MALE, income=45,000+, children=2} → {pep=YES}
    a. Support = 0.02
    b. Confidence = 1.0
    c. Lift = 2.19

3. {income=45,000+, married=NO, mortgage=NO} → {pep=YES}
    a. Support = 0.02
    b. Confidence = 1.0
    c. Lift = 2.19

4. {age=forties, region=TOWN, children=1} → {pep=YES}
    a. Support = 0.027
    b. Confidence = 1.0
    c. Lift = 2.19

5. {age=forties, income=15,000-24,999, children=1} → {pep=YES}
    a. Support = 0.03
    b. Confidence = 1.0
    c. Lift = 2.19

What can be inferred from the association rules above is that it is hard to pinpoint any one factor as a leading contributor to one enrolling in the PEP. What is commonly observed however is that those who tend to be 40 years of age or older and make $45k plus in annual income have a higher likelihood of enrolling in the PEP. This is an incredibly useful input to the bank when determining how to strategically target individuals who are more likely to enroll.

Conclusion

The analysis of association rules regarding financial services, specifically concerning the marketing, release, and adoption of the Personal Equity Plan, offers valuable insights into consumer behavior and preferences. Through examining demographic profiles and banking behaviors, financial institutions gain an understanding of the motivations guiding individuals' investment decisions and financial planning strategies. Personal finance can be described by evolving consumer needs and increasing accessibility to investment opportunities, which highlights the importance of tailored offerings and strategic approaches by financial institutions. As consumers preferences shift and access to a wide breath of resources becomes more prevalent, banks and advisors will need to adapt their product and service offerings to promote financial inclusion and empower consumers with the tools needed to navigate the complexities of personal finance confidently.

Through the application of the Apriori algorithm and association rule mining techniques, banks can discover meaningful patterns and associations within transactional data, shedding light on customer preferences and behaviors that inform decision-making strategies. The identification of key associations, such as individuals aged 40 and above and with higher incomes, shows the relevance of targeted marketing and product development efforts aimed at enhancing customer engagement and satisfaction.

The analysis of association rules not only enhances ones understanding of consumer behavior but also equips financial institutions with actionable insights to drive innovation, improve customer experiences, and promote financial well-being in an increasingly dynamic landscape. As banks continue to leverage data analytics and machine learning, the pursuit of financial inclusion and empowerment will remain fundamental to the mission of serving consumer needs effectively and honestly.