HW3 Association Rules

Rafal Rudzinski

Introduction

According to the U.S. Census Bureau, only 32% of Americans are investing their retirement. Even less Americans invest in the stock market outside of company 401k plans even as investing has become wildly accessible to retail consumers. As financial institution which offers investing plans such as the Personal Equity Plan (PEP), it is essential to understand how and why customers invest in order to offer the best products and maximize earnings.

An investment institution general profits from fees and commissions on underwriting securities. Therefore it is important for financial institutions to understand their clients so that they can strictly target one who are most likely to purchase their newest product. Being able to narrow down a specific demographic is key in maximizing profits and reducing marketing costs.

In the following analysis we examine bank data which include customer information. We will use this dataset to learn which customers are most likely to purchase a new PEP plan if it was offered at this financial institution. By associating customer's demographic and banking information we hope to learn which customers should be the focus of future marketing campaigns.

Analysis and Models

About the Data

Bank data was provided by DePaul University in csv format. It contains 600 observations of customer data including demographic and account information. *Table 1* lists all 12 features with short descriptions:

feature	**description**
id	A unique identification number
age	Age of customer in years
sex	Male/Female
region	Inner city/rural/suburban/town
income	Income of Customer*
married	Is the customer married (yes/no)
children	Number of children
car	Does the customer own a car (yes/no)
save_acct	Does the customer have a savings account (yes/no)
current_acct	Does the customer have a current account (yes/no)
mortgage	Does the customer have a mortgage (yes/no)
pep	Did the customer buy a PEP after the last mailing (yes/no)

Table 1

Reading the data

Blank spaces were replaced with NAs:

```
data <- read.csv("C:\\Users\\rafal\\OneDrive - Syracuse University\\IST 707\\Week 3\\bankdata_csv_all.cs
                 na.string = c(""))
```

Data Structure

Check data dimensions:

```
dim(data)
```

```
## [1] 600  12
```

Identify potential problems:

```
str(data)
```

```
## '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" ...
```

Verify dataset does not have missing values:

```
colSums(is.na(data))
```

```
##          id         age         sex      region      income     married
##           0           0           0           0           0           0
##    children         car    save_act current_act    mortgage         pep
##           0           0           0           0           0           0
```

Examining distribution of numeric features identifies age and income as possible candidates for discretization. Plotting these two features to visualize their distribution will help make the final determination.

There appears to be a fairly even distribution of age ranging from 18 to 67 years old. The median age of bank customers is 42 years old. We also notice that the mean and median are almost the same which confirms that the distribution has almost no skewness and is largely symmetrical.

The range of incomes is 5,014 to 63,130 with a median income of 27,524. This indicates a positive skew which we will address through discretization.

```
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   30.00   42.00   42.40   55.25   67.00
```

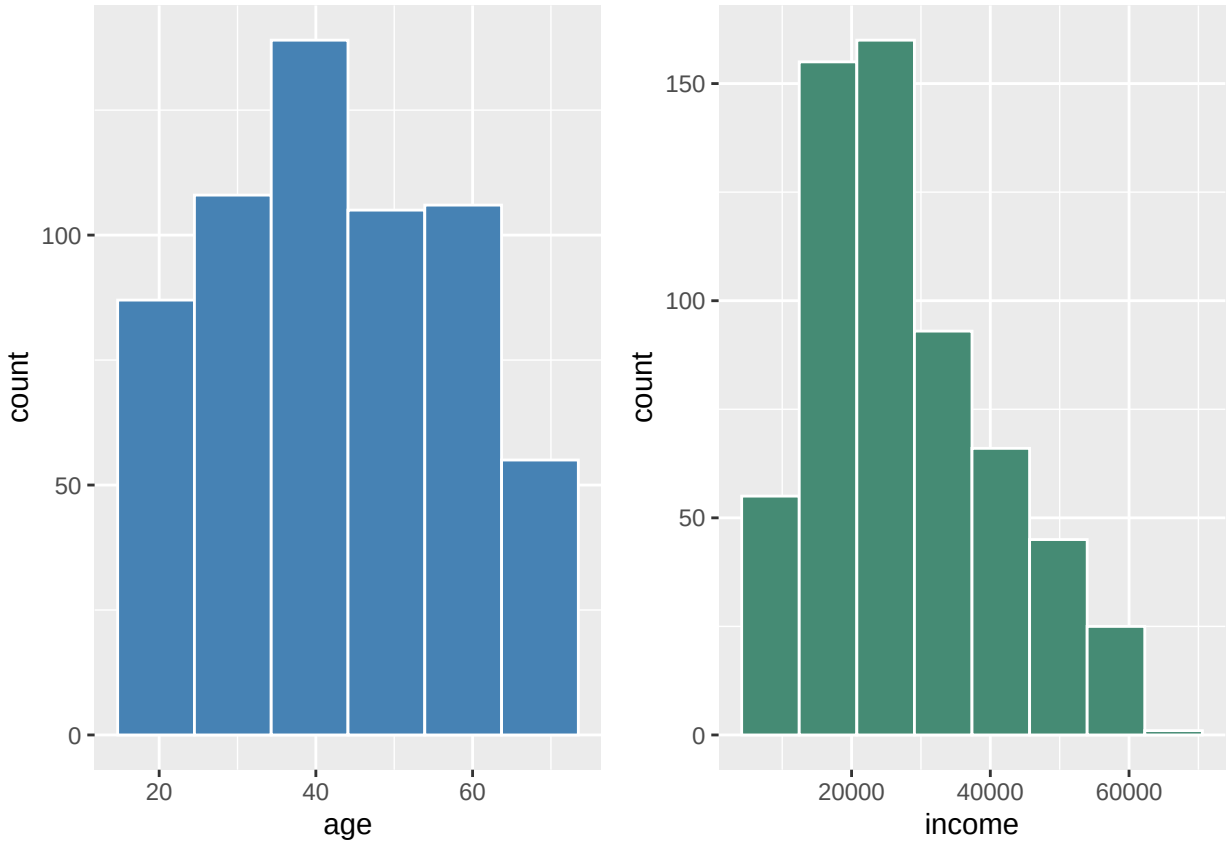
```
summary(data$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5014   17265   24925   27524   36173   63130
```

```
ageHist <- data %>%
  ggplot(aes(age)) +
  geom_histogram(bins = 6, fill="steelblue", col="white")

incomeHist <- data %>%
  ggplot(aes(income)) +
  geom_histogram(fill="aquamarine4", col="white", bins = 8)

grid.arrange(ageHist, incomeHist, nrow = 1)
```



Upon examining the data structure several issues were identified:

- ID feature can be removed, does not provide valuable information
- multiple numeric features require conversion to nominal or discretization

- multiple char features require conversion to nominal factor
- children feature requires conversion to ordinal factor
- verify uniqueness

Cleaning and Prep

We will fix all identified issues to prepare for further exploration and analysis.

```
#remove id field

#convert children to ordinal factor
data$children <- ordered(data$children)

#convert char features to factors
data$sex <- factor(data$sex)
data$region <- factor(data$region)
data$married <- factor(data$married)
data$car <- factor(data$car)
data$save_act <- factor(data$save_act)
data$current_act <- factor(data$current_act)
data$mortgage <- factor(data$mortgage)
data$pep <- factor(data$pep)

#discretize age and income
data$age <- cut(data$age, breaks = c(0,20,30,40,50,60,100),
                labels = c("teens", "twenties", "thirties", "forties", "fifties", "sixties"),
                right = FALSE)

data$income <- cut(data$income, breaks = c(0,15000,25000,35000,45000,100000),
                   labels = c("0-14999", "15,000-24,999", "25,000-34,999",
                              "35,000-44999", "45,000+"),
                   right = FALSE)
```

```
table(data$age)
```

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

```
table(data$income)
```

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

Data is now cleaned with no missing or duplicate values.

```
str(data)
```

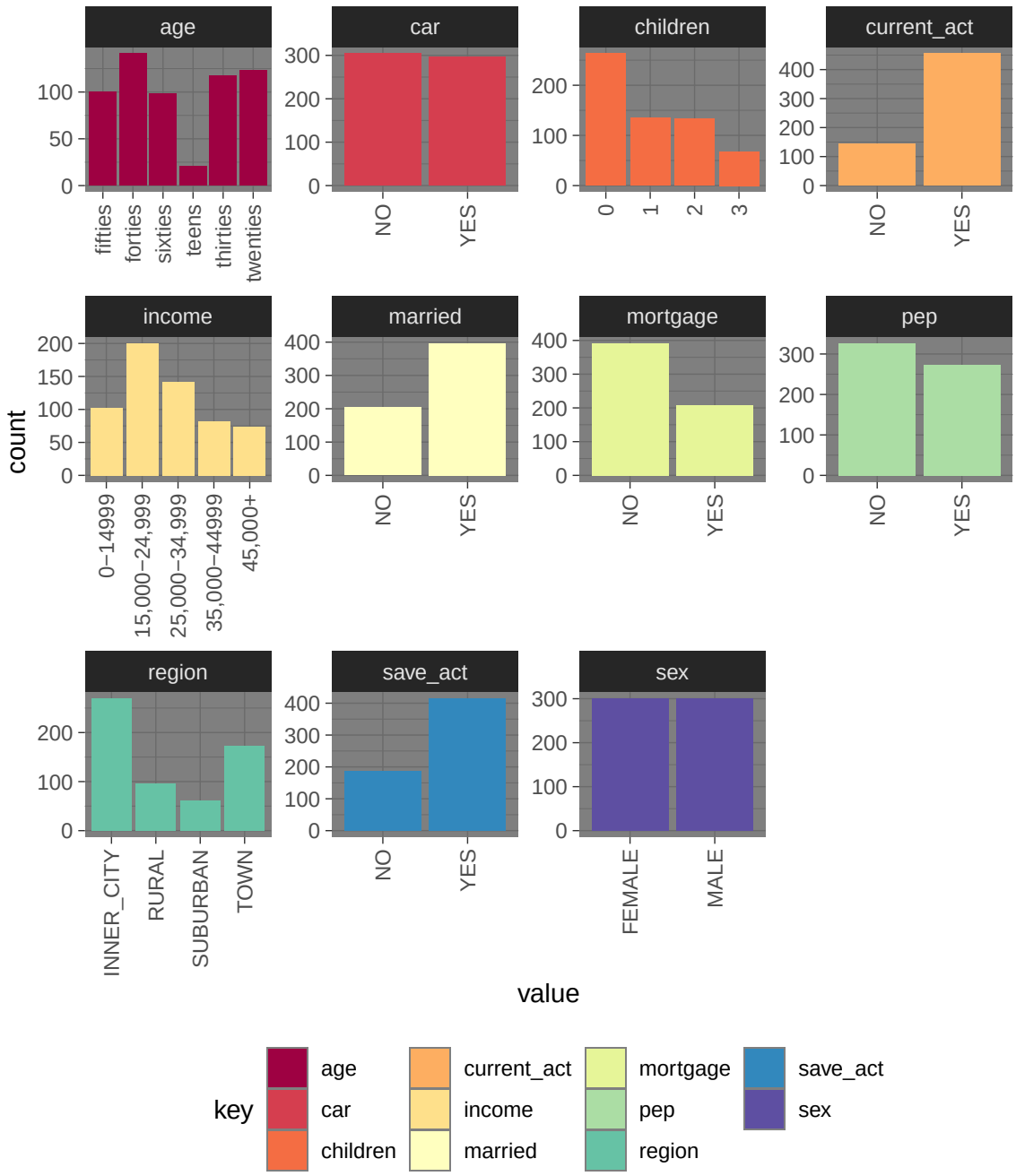
```
## 'data.frame':    600 obs. of  12 variables:
##  $ id         : chr  "ID12101" "ID12102" "ID12103" "ID12104" ...
```

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

Exploratory Data Analysis

Exploring the frequency distribution of all of the attributes reveals some insightful information about the customers.

- Almost half have no kids but are married and have a mortgage
- Most have a current account and a savings account
- Most have income less than 34,999
- Most reside in town or the inner city



Association Model (Apriori)

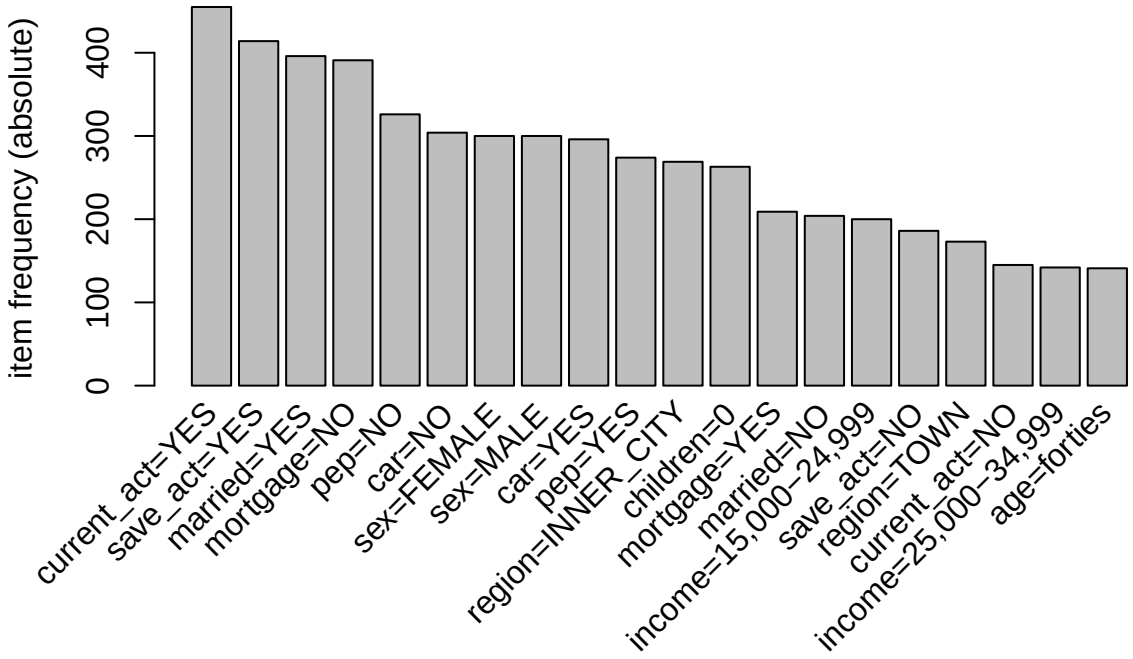
The Apriori algorithm allows us to generate association rules that can assist with understanding which customers are likely to obtain the new PEP (Personal Equity Plan). The Apriori algorithm uses 3 metrics to evaluate association rules:

1. Support - how popular an itemset is
2. Confidence - how often items A and B occur together
3. Lift - strength of a rule

```
tid <- as.character(data[["id"]])
data$id <- NULL
transactions <- as(data, "transactions")
transactionInfo(transactions)[["transactionID"]] <- tid
```

First, the most frequent items are identified in order to better understand what items are most likely to appear in mined rules and to prepare for making adjustments based on item frequency.

```
itemFrequencyPlot(transactions, topN=20, type="absolute")
```



Our initial attempt mined rules using the apropri algorithm set to support of 0.002 and confidence of 0.5.

The top 5 rules had a high confidence and lift values but very low support. Using the first rule as an example we can interpret the metrics as:

- A 0.003 support means that of all the transaction, only 0.03% represent the lhs and rhs combinations.
- A confidence of 1 indicates that of all transaction where the customer was lived in the suburbs, had an income between 0-14,9999, no more than 1 child, and no car they were 100% likely to also be a teenager.
- A lift of 28.57 indicated that a teenager is 28 times more likely to live in the suburbs, have an income of 0-14,999 and have 1 child and no car.

```
rules_pep <- apriori(transactions, parameter = list(supp = 0.002, conf = 0.5))
rules_pep <- sort(rules_pep, decreasing = TRUE, by="lift")
```

```
inspect(rules_pep[1:5])
```

```
##     lhs                   rhs             support confidence    coverage     lift count
## [1] {region=SUBURBAN,
##      income=0-14999,
##      children=1,
##      car=NO}          => {age=teens} 0.003333333          1 0.003333333 28.57143     2
## [2] {region=SUBURBAN,
##      children=1,
##      car=NO,
##      pep=NO}          => {age=teens} 0.003333333          1 0.003333333 28.57143     2
## [3] {sex=MALE,
##      region=RURAL,
##      income=0-14999,
##      car=YES}         => {age=teens} 0.003333333          1 0.003333333 28.57143     2
## [4] {sex=MALE,
##      region=RURAL,
##      income=0-14999,
##      save_act=YES}    => {age=teens} 0.003333333          1 0.003333333 28.57143     2
## [5] {income=0-14999,
##      car=YES,
##      mortgage=YES,
##      pep=YES}         => {age=teens} 0.003333333          1 0.003333333 28.57143     2
```

Because the support at 0.003 is very low, we made adjustments to the apriori algorithm to mine stronger rules. We found the strongest rules after setting support to 0.021 and confidence at 0.91. This configuration started to produce rules with high lift and relatively strong support.

```
inspect(rules_pep[1:5])
```

```
##     lhs                   rhs                  support confidence   coverage     lift count
## [1] {age=sixties,
##      car=YES,
##      save_act=YES,
##      mortgage=NO,
##      pep=YES}         => {income=45,000+} 0.03000000  0.9473684 0.03166667 7.681366    18
## [2] {age=sixties,
##      married=NO,
##      save_act=YES,
##      current_act=YES,
##      pep=YES}         => {income=45,000+} 0.02500000  0.9375000 0.02666667 7.601351    15
## [3] {age=sixties,
##      married=NO,
##      save_act=YES,
##      mortgage=NO,
##      pep=YES}         => {income=45,000+} 0.02333333  0.9333333 0.02500000 7.567568    14
## [4] {age=sixties,
##      car=YES,
##      save_act=YES,
##      current_act=YES,
##      mortgage=NO,
##      pep=YES}         => {income=45,000+} 0.02333333  0.9333333 0.02500000 7.567568    14
```

```
## [5] {age=sixties,
##      married=NO,
##      save_act=YES,
##      current_act=YES,
##      mortgage=NO,
##      pep=YES}         => {income=45,000+} 0.02166667  0.9285714 0.02333333 7.528958    13
```

Our goal is to explore associations related to the "PEP" attribute to determine what type of customers are more likely to obtain the new PEP (Personal Equity Plan). To do this we will set the right hand side of the rule to PEP.

Based on the 5 most interesting rules we present out recommendations for the type of clients that are likely to want to obtain the new PEP:

1. *{married=NO,children=0,save_act=YES,mortgage=NO} => {pep=YES}*
    - support: 0.06
    - confidence: 1.0
    - lift: 2.19
    - This is an interesting pattern because this seems to reflect on customers most likely are new to investing or may not have much equity yet.
    - We recommend targeting markets with a younger professional demographic such as the tech industry. They may have the disposable income and desire to invest.
2. *{sex=MALE,married=NO,children=0,save_act=YES,mortgage=NO} => {pep=YES}*
    - support: 0.03
    - confidence: 1.0
    - lift: 2.19
    - Unmarried males with no family or mortgage may have disposable income to invest.
    - Target advertisements at sporting events, bars.
3. *{age=sixties,region=SUBURBAN,income=45,000+} => {pep=YES}*
    - support: 0.013
    - confidence: 1.0
    - lift: 2.19
    - This demographic includes customers who are getting ready to retire.
    - Sending mail advertisements or advertising in newspapers could reach more customers in this demographic.
4. *{region=SUBURBAN,income=45,000+,children=2} => {pep=YES}*
    - support: 0.01
    - confidence: 1.0
    - lift: 2.19
    - Suburban middle class families are often interested in investing for retirement, their children.
    - Sponsoring a kisa soccer or baseball team could provide more exposure.
5. *{children=1,save_act=YES,current_act=YES} => {pep=YES}*
    - support: 0.105
    - confidence: 0.86
    - lift: 1.89
    - Current customers in good standing with an active savings account who are also parents may be thinking long term about investing in order to set up their child for success.
    - Email and notification advertisements on the bank web portal could increase their exposure to the new PEP.

Conclusion/Results

Analysis of customer banking information and demographics data resulted in some insightful association rules that should provide a better understanding of the type of customer that is most likely to have a favorable response to a PEP offer by the bank. To summarize our findings we will explain rule 5: {children=1,save_act=YES,current_act=YES} => {pep=YES} in more detail to provide a more detailed explanation fo the associated metrics.

Rule 5 identifies an customer who is currently banking at the institution and has an active savings account. They also have 1 child. A support value of 0.105 indicates that of all 600 transactions, that the items in rule 5 appear together roughly 10% of the time. That value might appear to be somewhat low but in this case we are not necessarily concerned as much with how customer frequency.

In fact we are far more concerned with the confidence metric which tells us how often does the customer in rule 5 actually purchase a PEP plan. In the case or rule 5, a customer with those metrics is expected to purchase a PEP plan 86% of the time. That is extremely important from a marketing standpoint.

Finally, a lift of 1.89 indicates a strong relationship between the customer and the likelihood they will purchase the PEP plan. Another word, this customer has a positive effect on the acceptance of a PEP plan if offered. Armed with these 5 rules and the interpretation of their metrics, we are confident that a target marketing campaign of a new PEP plan will be succesfull.