Code

Kevin_Pollard_HW3

Kevin Pollard

4/22/2020

Introduction

The marketing department of KNJ Financial keeps records on customers, including demographic information and, number of type of accounts. When launching a new product, such as a “Personal Equity Plan” (PEP), a direct mail piece, advertising the product, is sent to existing customers, and a record kept as to whether that customer responded and bought the product. Based on this store of prior experience, the managers decide to use data mining techniques to build customer profile models.

The personal equity plan was designed to encourage investment by individuals. Many plans required a minimum amount to be invested, depending on the type of plan and the plan manager’s requirements. The individuals who responded to the PEP direct mailer piece indicate a wiliness to both save money and respond marketing mailers from KNJ Financial. It is a fair assumption these same customer will likely repond to future campaigns, but more analysis is required.

The type of analysis recommended for the managers to leverage is called **propensity to buy.** The purpose of a propensity to buy analysis is to understand the likelihood a customer will be predisposed to purchasing a product based upon purchases they’ve already made. A few data mining techniques will be introduced in the sections below, and concepts will be explained along the way.

Analysis and Models

About the Data

Data Loading

KNJ finance has provided the bankdata_csv_all.csv file to support the analysis.

Data Structure

The data has 600 observations with 12 features to describe each. The id field has no value and will be elinated from the analysis.

The data contains of a number of the following fields:

- 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 saving 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

```
## 'data.frame':    600 obs. of  12 variables:
##  $ id         : Factor w/ 600 levels "ID12101","ID12102",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ age        : int  48 40 51 23 57 57 22 58 37 54 ...
##  $ 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     : num  17546 30085 16575 20375 50576 ...
##  $ married    : Factor w/ 2 levels "NO","YES": 1 2 2 2 2 2 1 2 2 2 ...
##  $ children   : int  1 3 0 3 0 2 0 0 2 2 ...
##  $ 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 ...
```

Data Tidiness and Completeness

The banking data is clean with no NA, incomplete sets, or duplicate records.

```
##                                              a
## 1 Complete Cases?  TRUE  Incomplete Count:  0
```

columns	percent_missing
id	0
age	0
sex	0
region	0
income	0
married	0
children	0
car	0
save_act	0
current_act	0
mortgage	0
pep	0

columns	count_missing
id	0

columns	count_missing
age	0
sex	0
region	0
income	0
married	0
children	0
car	0
save_act	0
current_act	0
mortgage	0
pep	0

d

Duplicate Count: 0

```
## [1] "summary"
##   pep           income            age          married
##  NO :326   Min.   : 5014   Min.   :18.00   NO :204
##  YES:274   1st Qu.:17265   1st Qu.:30.00   YES:396
##            Median :24925   Median :42.00
##            Mean   :27524   Mean   :42.40
##            3rd Qu.:36173   3rd Qu.:55.25
##            Max.   :63130   Max.   :67.00
## [1] "summary"
##     children     current_act save_act   car
##  Min.   :0.000   NO :145     NO :186   NO :304
```

```
##  1st Qu.:0.000   YES:455     YES:414   YES:296
##  Median :1.000
##  Mean   :1.012
##  3rd Qu.:2.000
##  Max.   :3.000
```

Proportions of the feature data are found below here. The Pep proportions of YES = 0.46 means a very large portion of KNJ customers responded to the pep mailer. This bodes well, since past responders or purchasers of services tend to respond again.

Checking proportions of the original data.

- "Pep proportions NO= 0.54 YES= 0.46"
- "Car proportions NO= 0.51 YES= 0.49"
- "Married proportions NO= 0.34 YES= 0.66"
- "Children proportions 0= 0.44 1= 0.22 2= 0.22 3= 0.11"
- "Mortgage proportions NO= 0.65 YES= 0.35"
- "Savings Account proportions NO= 0.31 YES= 0.69"

The table below shows top 10 customers arranged by income **income high to low.**.

- Most of the top 10 by income have responded to the Personal Equity Plan" mailer.
- All have savings and current accounts.
- The majority are married females who are retirement age.
- This cross section doesn't have a lotof mortgages, so maybe the mortagees are paid off and kids left the nest.

These points should help inform a deeper analysis.

id	agesex	region	incomemarried	childrencar	save_act	current_act	mortgage	pep
ID12291	67FEMALE	SUBURBAN	63130.1YES	2YES	YES	YES	NO	YES
ID12558	65FEMALE	INNER_CITY	61554.6YES	0NO	YES	YES	NO	NO
ID12371	67FEMALE	SUBURBAN	60747.5NO	2NO	YES	YES	YES	YES
ID12605	63FEMALE	INNER_CITY	59805.6YES	1YES	YES	YES	NO	YES
ID12111	66FEMALE	TOWN	59803.9YES	0NO	YES	YES	NO	NO
ID12235	66FEMALE	TOWN	59503.8YES	2YES	YES	YES	YES	YES
ID12307	63MALE	INNER_CITY	59409.1NO	0YES	YES	YES	NO	YES
ID12631	64MALE	SUBURBAN	59175.1YES	1NO	YES	YES	NO	YES
ID12532	64FEMALE	TOWN	58367.3YES	1YES	YES	YES	NO	YES
ID12513	67FEMALE	INNER_CITY	58092.0NO	2YES	YES	YES	NO	YES

The table below shows top 10 customers arranged by **income low to high.**

- Most did not respond to the pep mailer

- The majority are not married

These points may also be significant.

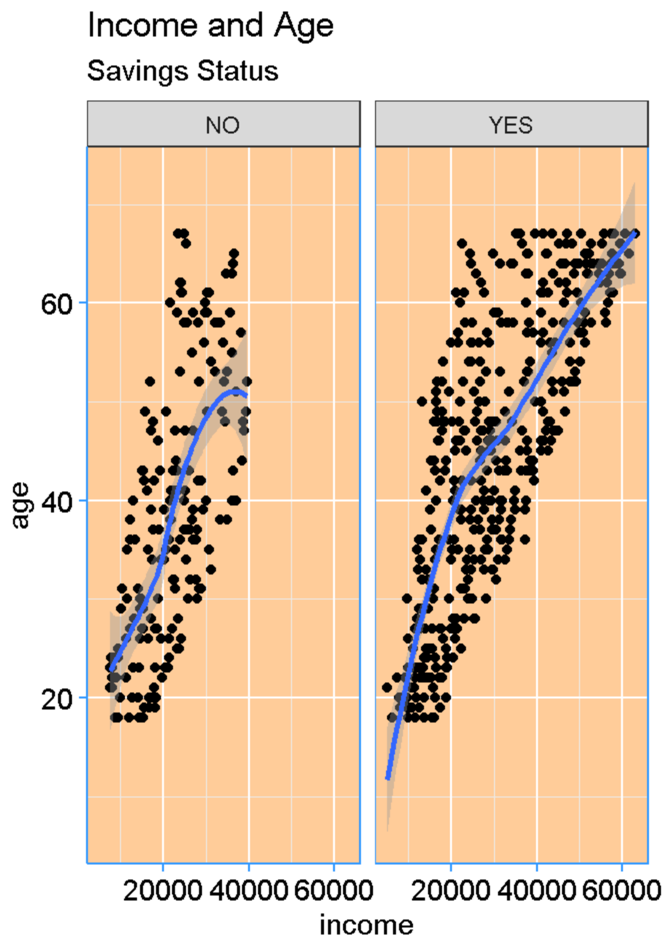
id	age	sex	region	income	married	children	car	save_act	current_act	mortgag
ID12165	21	MALE	TOWN	5014.21	NO	0	YES	YES	YES	YES
ID12679	18	MALE	INNER_CITY	6294.21	NO	0	NO	YES	YES	YES
ID12512	22	MALE	INNER_CITY	7304.20	NO	0	YES	YES	YES	YES
ID12264	21	FEMALE	TOWN	7549.38	NO	1	YES	NO	YES	NO
ID12347	23	FEMALE	INNER_CITY	7606.25	NO	3	YES	NO	NO	NO
ID12656	20	FEMALE	TOWN	7723.93	YES	2	YES	YES	YES	NO
ID12473	24	FEMALE	INNER_CITY	7756.36	NO	0	NO	NO	NO	NO
ID12197	22	MALE	INNER_CITY	7948.62	YES	1	NO	NO	NO	YES
ID12651	23	FEMALE	INNER_CITY	8020.19	YES	1	YES	NO	YES	NO
ID12172	21	MALE	INNER_CITY	8062.73	NO	0	NO	NO	YES	NO

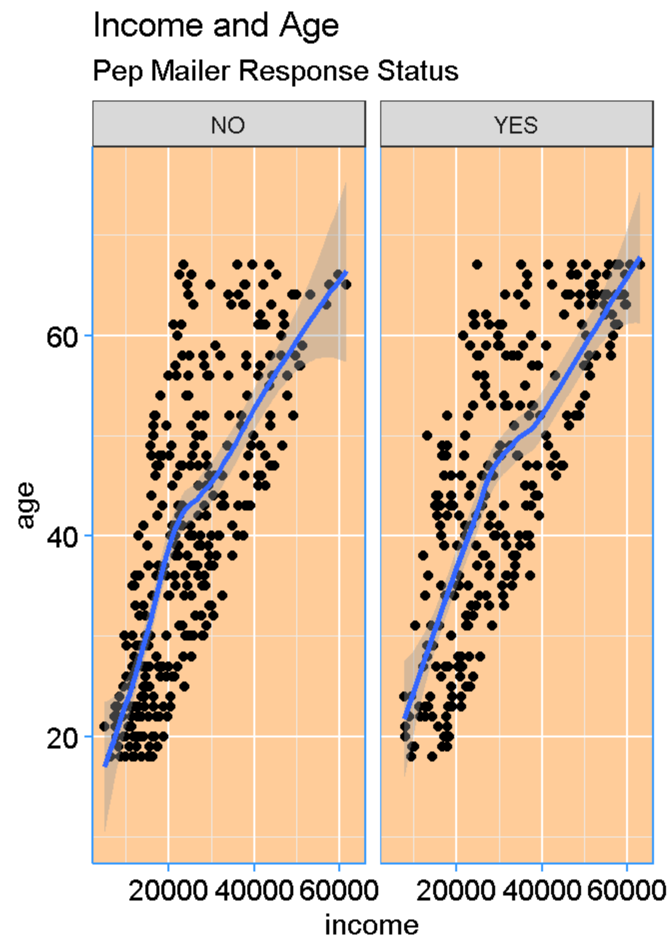
Summary Stats Analysis

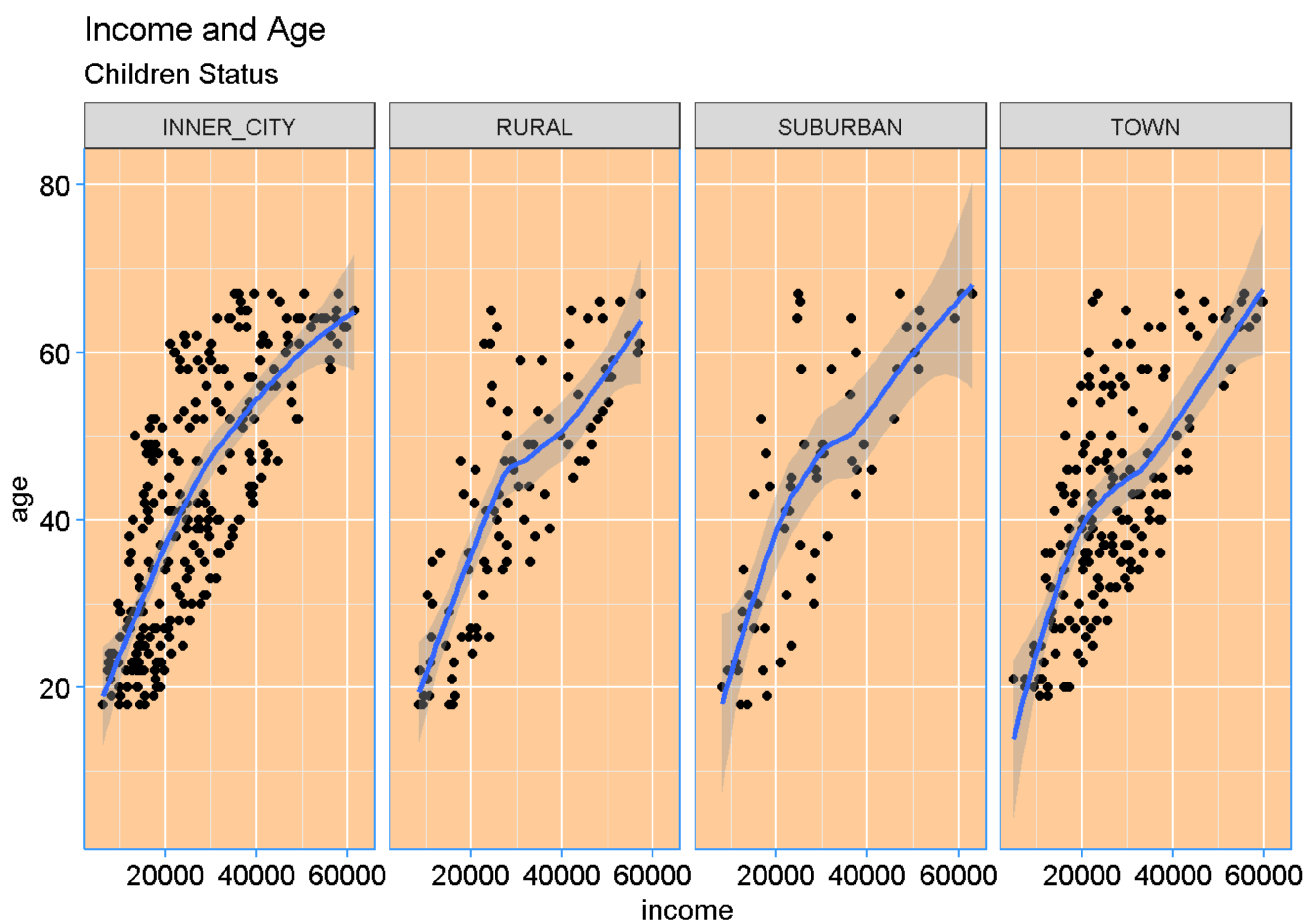
- The average age of the customer population is 42, and 44% of them have no children.
- Average income is $27,524.

```
## [1] 600  12
## [1] "---------------------------------------------------------------"
## [1] "age"
## [1] "---------------------------------------------------------------"
## [1] "mean:  42.395"
## [1] "median:  42"
## [1] "min:  18"
## [1] "max:  67"
```

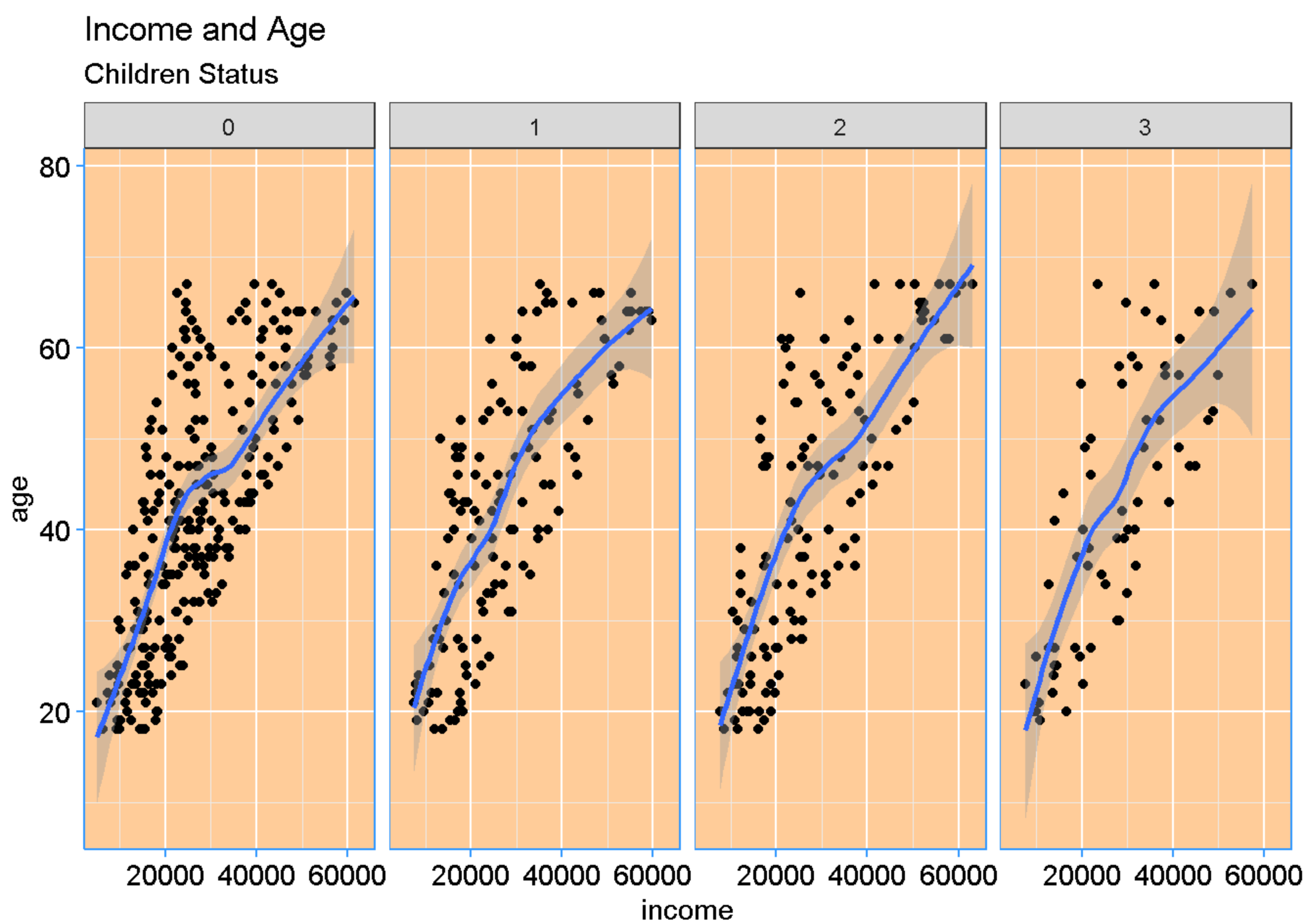
```
## [1] "range:  49"
## [1] "sd:  14.424947377538"
## [1] "quantile:  18"    "quantile:  30"    "quantile:  42"    "quantile:
55.25"
## [5] "quantile:  67"
## [1] "IQR:  25.25"
## [1] "----------------------------------------------------------------"
## [1] "income"
## [1] "----------------------------------------------------------------"
## [1] "mean:  27524.0312166667"
## [1] "median:  24925.3"
## [1] "min:  5014.21"
## [1] "max:  63130.1"
## [1] "range:  58115.89"
## [1] "sd:  12899.4682456306"
## [1] "quantile:  5014.21"   "quantile:  17264.5"   "quantile:  24925.3"
## [4] "quantile:  36172.675" "quantile:  63130.1"
## [1] "IQR:  18908.175"
## [1] "----------------------------------------------------------------"
## [1] "children"
## [1] "----------------------------------------------------------------"
## [1] "mean:  1.01166666666667"
## [1] "median:  1"
## [1] "min:  0"
## [1] "max:  3"
## [1] "range:  3"
## [1] "sd:  1.05675214567302"
## [1] "quantile:  0" "quantile:  0" "quantile:  1" "quantile:  2" "quanti
le:  3"
## [1] "IQR:  2"
```


Income and Age
Savings Status
NO
YES
age
60
40
20
20000 40000 60000
20000 40000 60000
income

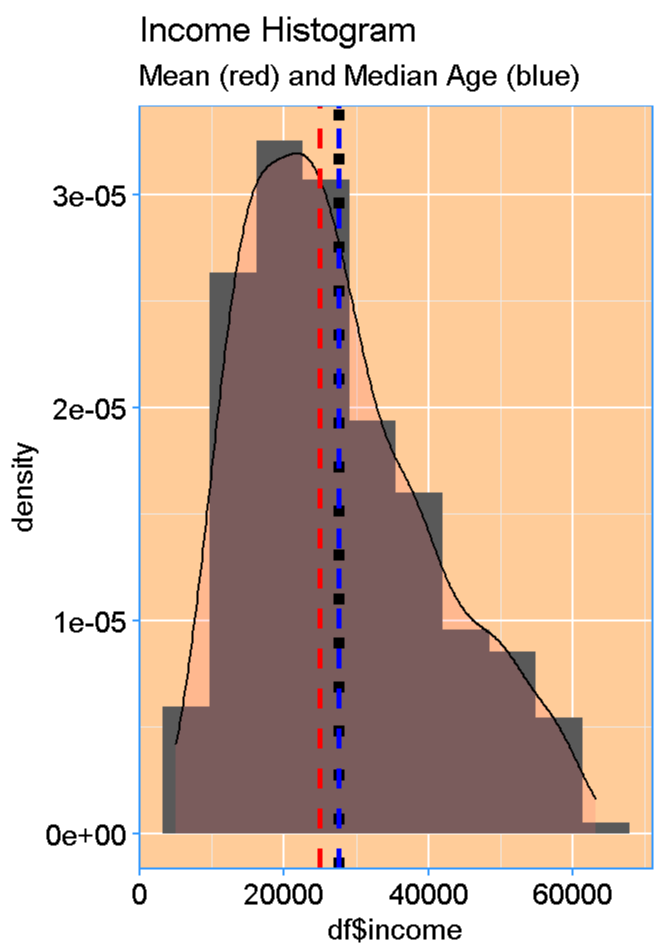

Income and Age
Pep Mailer Response Status
NO
YES
age
60
40
20
20000 40000 60000
20000 40000 60000
income

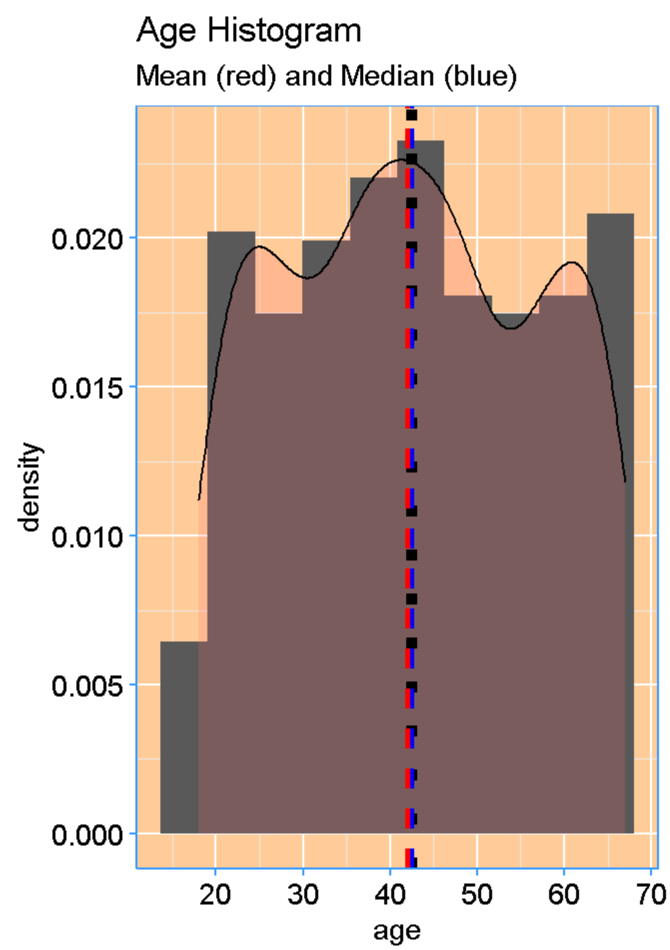


Income and Age
Children Status
INNER_CITY
RURAL
SUBURBAN
TOWN
age
80
60
40
20
20000
40000
60000
income



Income and Age
Children Status
0
1
2
3
80
60
40
20
age
20000 40000 60000
income





Data Transforms

Discretization and numeric-to-nominal transformation is necessary for the Apriori algorithm to be used. The following transformations will be applied to the data.

- Discretize age by customized bin
- Discretize income by equal-width bin
- Convert numeric to nominal for “children”
- Now the second step of conversion, changing “YES” to “[variable_name]=YES”.

```
## 'data.frame':    600 obs. of  11 variables:
##  $ age         : int  48 40 51 23 57 57 22 58 37 54 ...
##  $ 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     : num  17546 30085 16575 20375 50576 ...
## $ married    : Factor w/ 2 levels "NO","YES": 1 2 2 2 2 2 1 2 2 2 ...
## $ children   : int  1 3 0 3 0 2 0 0 2 2 ...
## $ 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 ...
```

age	sex	region	income	married	children	car	save_act	current_act	mortgage	pe
fourties	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO	NO	NO	YE
thirties	MALE	TOWN	30085.1	YES	3	YES	NO	YES	YES	NO
fifties	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES	YES	NO	NO
twenties	FEMALE	TOWN	20375.4	YES	3	NO	NO	YES	NO	NO
fifties	FEMALE	RURAL	50576.3	YES	0	NO	YES	NO	NO	NO
fifties	FEMALE	TOWN	37869.6	YES	2	NO	YES	YES	NO	YE

age	sex	region	income	married	children	car	save_act	current_act	
fourties	FEMALE	INNER_CITY	(5.01e+03,2.44e+04]	NO	1	NO	NO	NO	N
thirties	MALE	TOWN	(2.44e+04,4.38e+04]	YES	3	YES	NO	YES	Y
fifties	FEMALE	INNER_CITY	(5.01e+03,2.44e+04]	YES	0	YES	YES	YES	N
twenties	FEMALE	TOWN	(5.01e+03,2.44e+04]	YES	3	NO	NO	YES	N
fifties	FEMALE	RURAL	(4.38e+04,6.31e+04]	YES	0	NO	YES	NO	N
fifties	FEMALE	TOWN	(2.44e+04,4.38e+04]	YES	2	NO	YES	YES	N

age	sex	region	income	married	children	car	save_act	current_act	r
fourties	FEMALE	INNER_CITY	(5.01e+03,2.44e+04]	NO	1	NO	NO	NO	N
thirties	MALE	TOWN	(2.44e+04,4.38e+04]	YES	3	YES	NO	YES	Y
fifties	FEMALE	INNER_CITY	(5.01e+03,2.44e+04]	YES	0	YES	YES	YES	N
twenties	FEMALE	TOWN	(5.01e+03,2.44e+04]	YES	3	NO	NO	YES	N
fifties	FEMALE	RURAL	(4.38e+04,6.31e+04]	YES	0	NO	YES	NO	N
fifties	FEMALE	TOWN	(2.44e+04,4.38e+04]	YES	2	NO	YES	YES	N

age	sex	region	income	married	children	car	save_act	cu
fourties	FEMALE	INNER_CITY	(5.01e+03,2.44e+04]	married=NO	1	car=NO	save_act=NO	cu
thirties	MALE	TOWN	(2.44e+04,4.38e+04]	married=YES	3	car=YES	save_act=NO	cu
fifties	FEMALE	INNER_CITY	(5.01e+03,2.44e+04]	married=YES	0	car=YES	save_act=YES	cu
twenties	FEMALE	TOWN	(5.01e+03,2.44e+04]	married=YES	3	car=NO	save_act=NO	cu
fifties	FEMALE	RURAL	(4.38e+04,6.31e+04]	married=YES	0	car=NO	save_act=YES	cu
fifties	FEMALE	TOWN	(2.44e+04,4.38e+04]	married=YES	2	car=NO	save_act=YES	cu

Application of the Apriori algothim

The **Apriori algorithm** can now be used on the transofrmed data to generate association rules. Parameters of support,confidence and lift will be leveraged to study rules that will inform KNJ’s targeted marking campaigns.

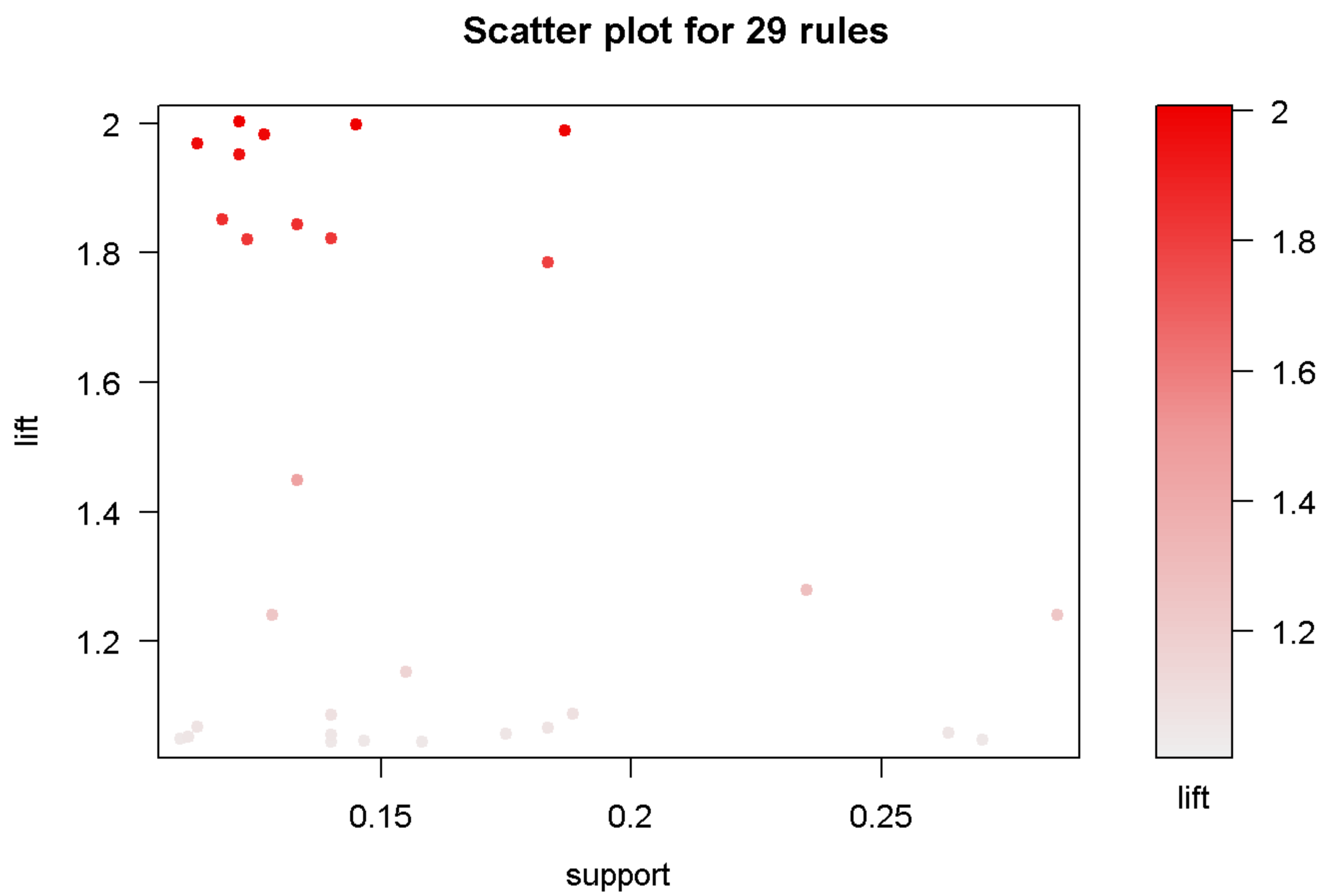
- **Support**: This measure gives an idea of how frequent an itemset is in all the transactions.
- **Confidence**: Confidence is an indication of how often the rule has been found to be true.
- **Lift**: Lift is the ratio of the observed support to that expected

Several different combinations of the support, and confidence were use to filter the universe of strong rules for the KNJ Financial data. * Set support: .11 * Set confidence: .79 * Set lift: >= 1

A set of 29 strong rules was obtained from these parameters about. Those 29 are filtered more to get the top 5 strong rules.

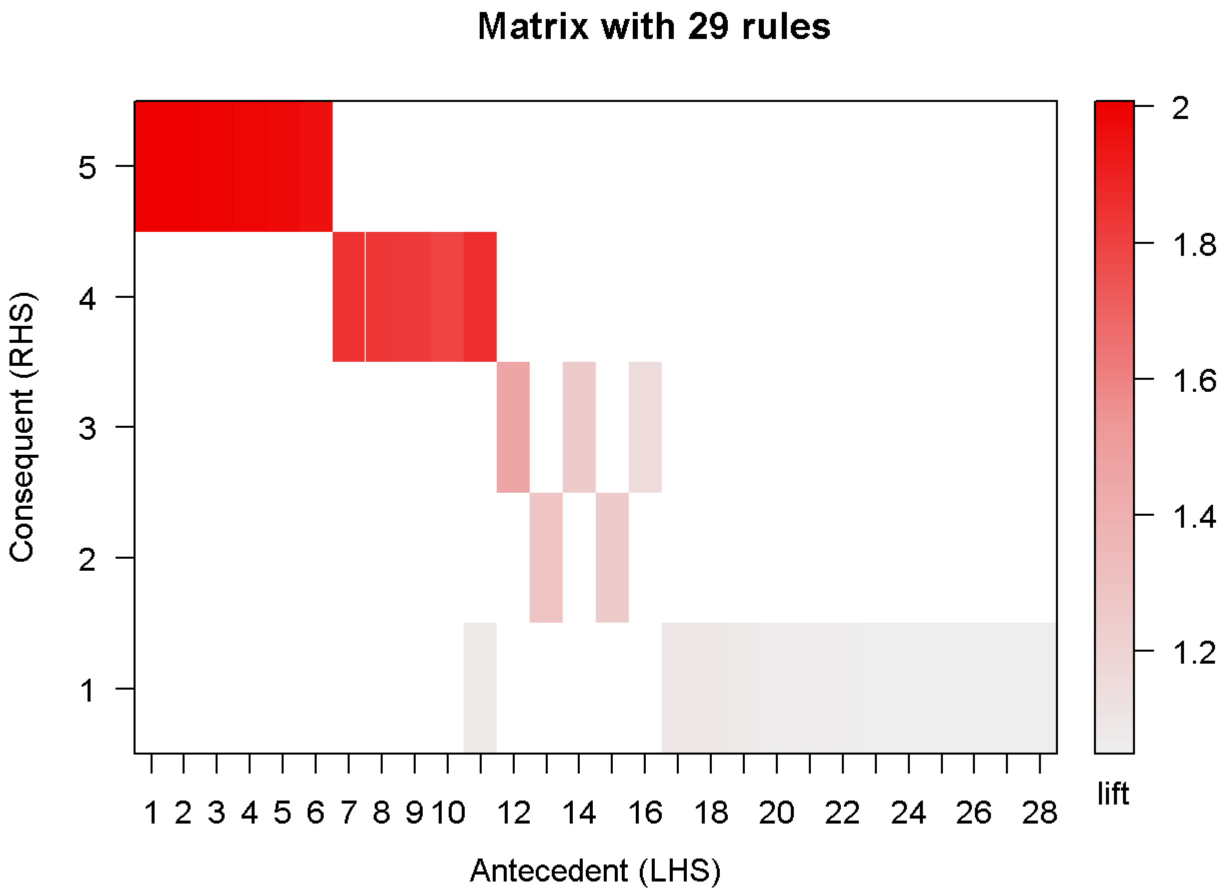
```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minl
en
##        0.75    0.1    1 none FALSE            TRUE       5    0.07
1
##  maxlen target   ext
##       3  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 42
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[31 item(s), 600 transaction(s)] done [0.00s].
## sorting and recoding items ... [30 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [194 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

Plot 29 Strong Rules




```
## Itemsets in Antecedent (LHS)
##  [1] "{age=twenties,married=married=YES}"
##  [2] "{age=twenties,current_act=current_act=YES}"
##  [3] "{age=twenties}"
##  [4] "{age=twenties,mortgage=mortgage=NO}"
##  [5] "{age=twenties,save_act=save_act=YES}"
##  [6] "{age=twenties,pep=pep=NO}"
##  [7] "{children=1,save_act=save_act=YES}"
##  [8] "{children=1,current_act=current_act=YES}"
##  [9] "{married=married=YES,children=1}"
## [10] "{children=1}"
## [11] "{children=1,mortgage=mortgage=NO}"
## [12] "{income=(4.38e+04,6.31e+04]}"
```

```
## [13] "{children=0,pep=pep=NO}"
## [14] "{age=old}"
## [15] "{mortgage=mortgage=NO,pep=pep=NO}"
## [16] "{mortgage=mortgage=YES,pep=pep=NO}"
## [17] "{married=married=NO,save_act=save_act=YES}"
## [18] "{married=married=NO,car=car=NO}"
## [19] "{car=car=NO,pep=pep=YES}"
## [20] "{car=car=NO,mortgage=mortgage=NO}"
## [21] "{sex=FEMALE,region=INNER_CITY}"
## [22] "{sex=FEMALE,married=married=NO}"
## [23] "{married=married=NO,pep=pep=NO}"
## [24] "{married=married=NO,children=0}"
## [25] "{married=married=NO}"
## [26] "{income=(2.44e+04,4.38e+04],pep=pep=YES}"
## [27] "{income=(2.44e+04,4.38e+04],car=car=NO}"
## [28] "{married=married=NO,pep=pep=YES}"
## Itemsets in Consequent (RHS)
## [1] "{current_act=current_act=YES}" "{married=married=YES}"
## [3] "{save_act=save_act=YES}"       "{pep=pep=YES}"
## [5] "{income=(5.01e+03,2.44e+04]}"
```



Rules sorted by support

This sort gives us perspective in terms of what high support means in the context of the KNJ Financial dataset. The highest support is .35 but that number tapers off very fast to .21 after just ten entries.

Rules sorted by confidence

Rules sorted by lift

Top 5 Rules

Sticking with the premise that past responders will have a tendency to respond again to future KNJ's mailers/campaigns, we use the left hand side (LHS) filter to specify pep=YES. Recall that **"pep=YES"** means the customer responded to the KNJ pepe mailer.

This results in only 5 rules, but they are very strong ones. Not only do they have good support, confidence and lift, but they also occur a lot. Recall there are only 600 observations in the KNJ dataset, so 110 occurances of a rule is high.

customer segment: Married couples with only 1 child who do business with KNJ, but not on the mortgage side.

- Children 1 rule with support of .183 is in the top 10 of all rules.
- Children 1 shows up in the antecedent of every rule in the top 5
- Current and Savings Account are YES
- No Mortgage though
- Married

Top Five Rules

1.) {children=1} {pep=pep=YES} 0.183 0.815 1.784 **110.000**

2.) {children=1,mortgage=mortgage=NO} {pep=pep=YES} 0.118 0.845 1.851 **71.000**

3.) {married=married=YES,children=1} {pep=pep=YES} 0.123 0.831 1.821 **74.000**

4.) {children=1,save_act=save_act=YES} {pep=pep=YES} 0.133 0.842 1.844 **80.000**

5.) {children=1,current_act=current_act=YES} {pep=pep=YES} 0.140 0.832 1.821 **84.000**

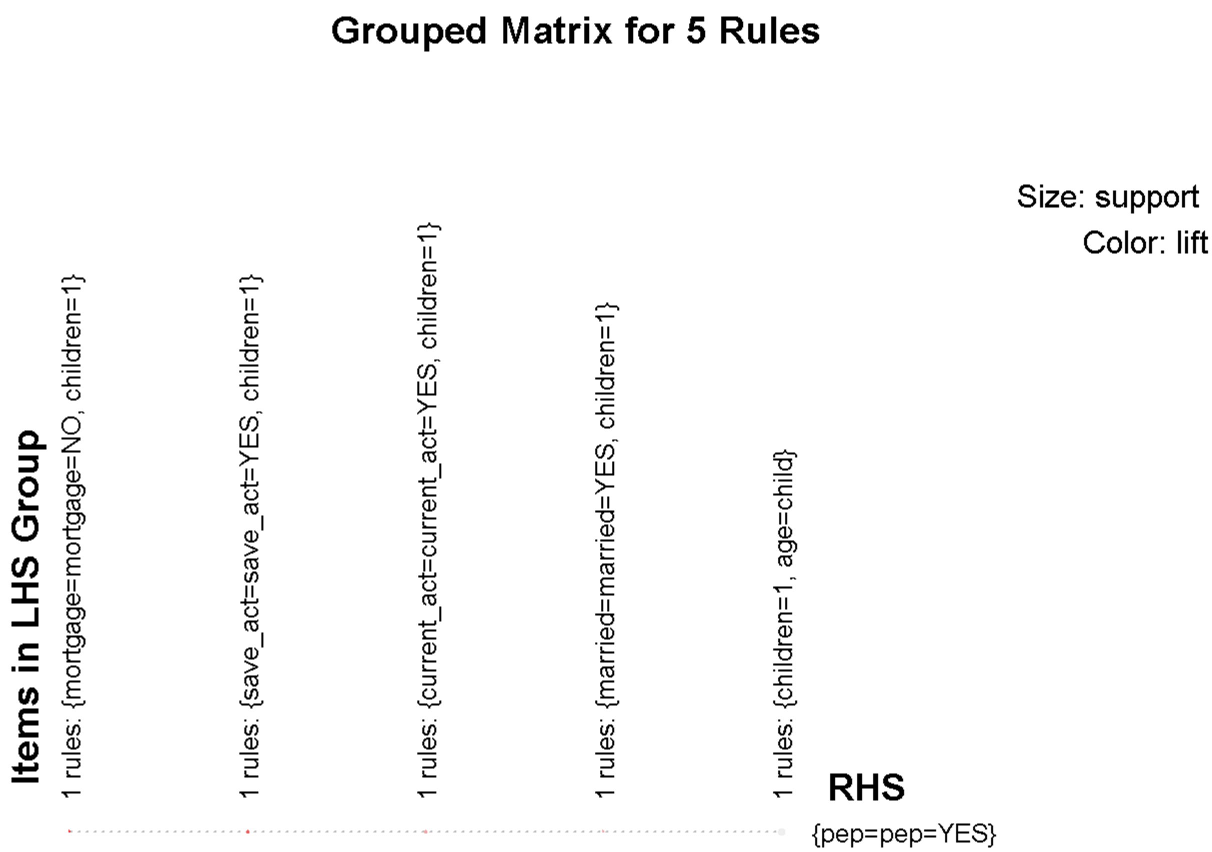
Data Resulting fromthe top 5 rules

The campaign developed below takes all five rules into account for maximum response. The mailer would be for 29 accounts. If KNJ would like to go broader adjustments to the rule filters will open up the list more. For example, you could have a broader campaign of 100 accounts by only including rule Children=1 combined with past PEP responders.

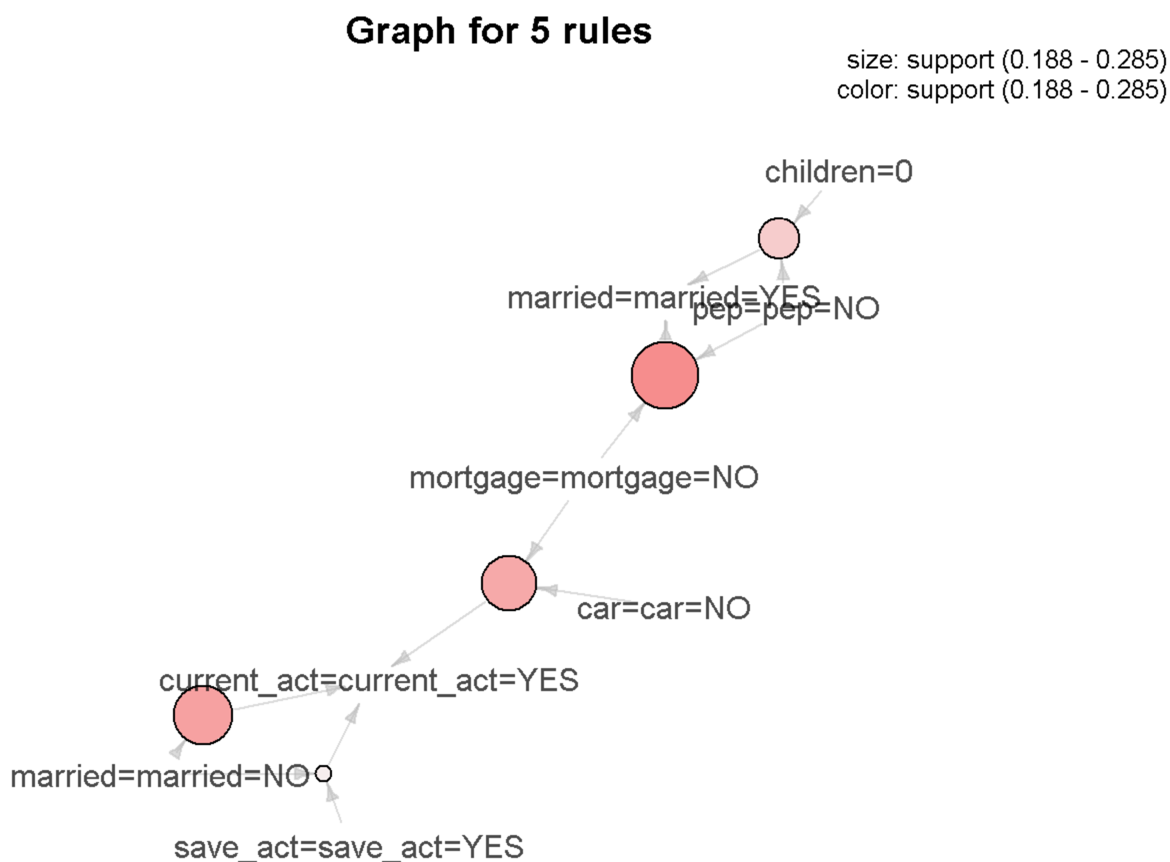
ArulViz Visualizations

Plotting Top 5 Rules

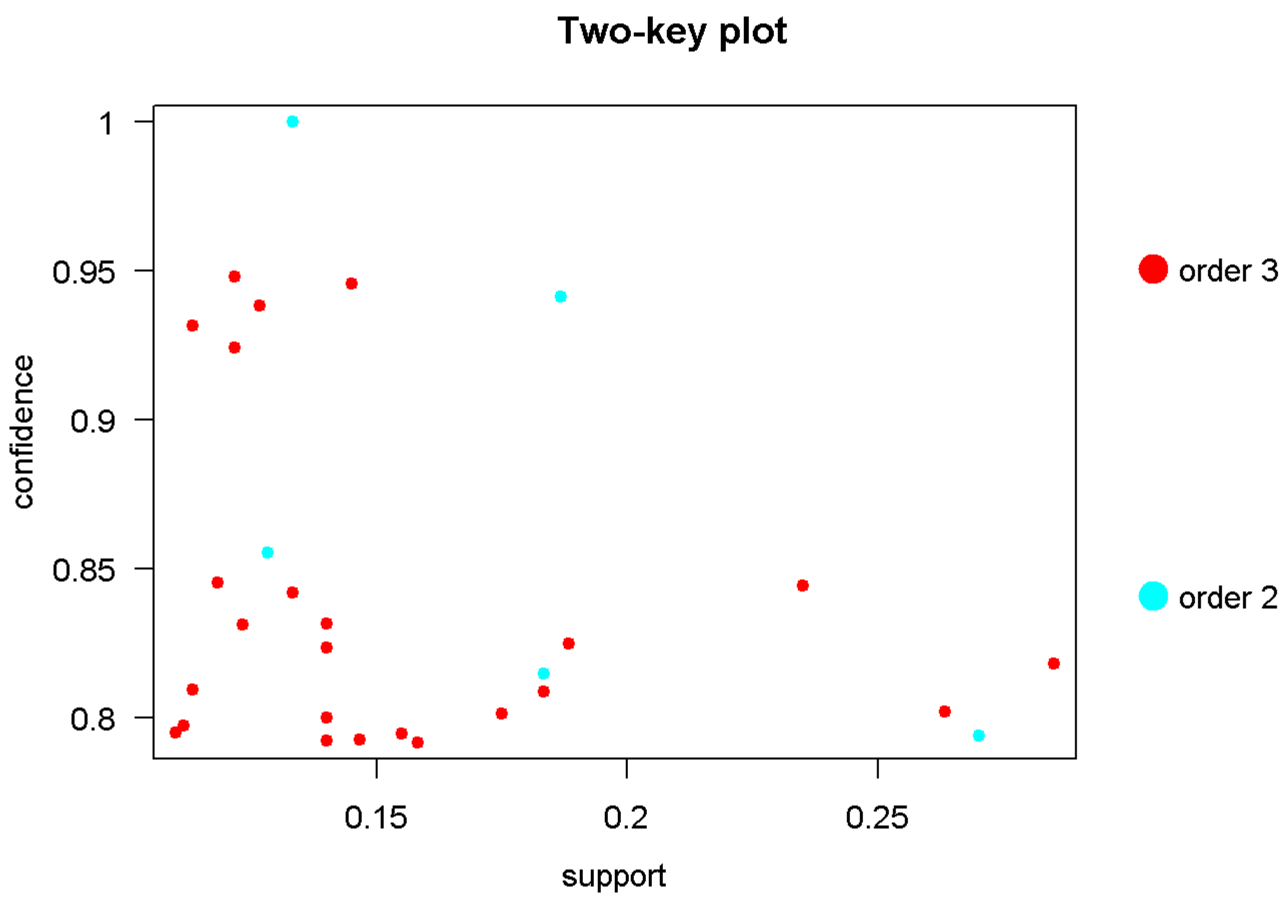
The plotting shows the lift comes mostly from having a savings and current account. No mortgage may indicate they have more disposable income.



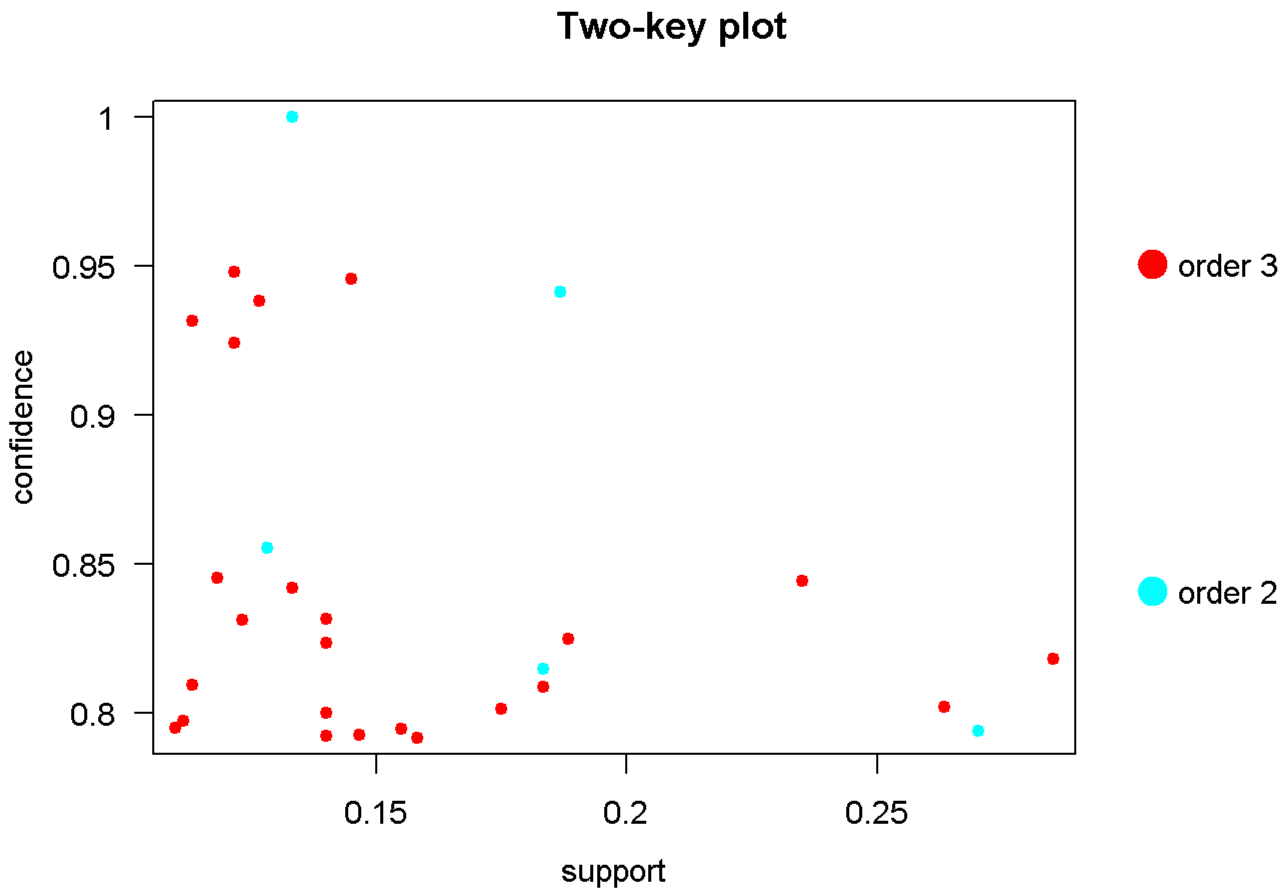
Plot Support



Plot Confidence


Two-key plot
1
0.95
0.9
0.85
0.8
confidence
0.15
0.2
0.25
support
order 3
order 2

Plot Lift



Models

Classification Models

Model Data Splitting

The data will be split into 2/3 for training set, and 1/3 for the testing set.

dimension_training_train	dimension_testing_train
400	200
12	12

	agesex	region	incomemarriedchildrencar	save_act	current_act	mortgagepep
76	59FEMALERURAL		35611NO	2YESNO	NO	NO YES

	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pe
298	18	MALE	RURAL	8639	YES	2	NO	NO	NO	NO	NO
414	26	FEMALE	INNER_CITY	16519	YES	0	YES	NO	YES	NO	YE
473	61	MALE	INNER_CITY	21140	YES	2	YES	YES	NO	NO	NO
325	36	MALE	SUBURBAN	28495	YES	0	NO	YES	YES	NO	NO
499	51	FEMALE	TOWN	43800	NO	0	NO	YES	YES	YES	NO

	id	age	sex	region	income	married	children	car	save_act	current_act	mort
76	ID12176	59	FEMALE	RURAL	35611	NO	2	YES	NO	NO	NO
298	ID12398	18	MALE	RURAL	8639	YES	2	NO	NO	NO	NO
414	ID12514	26	FEMALE	INNER_CITY	16519	YES	0	YES	NO	YES	NO
473	ID12573	61	MALE	INNER_CITY	21140	YES	2	YES	YES	NO	NO
325	ID12425	36	MALE	SUBURBAN	28495	YES	0	NO	YES	YES	NO
499	ID12599	51	FEMALE	TOWN	43800	NO	0	NO	YES	YES	YES

dimension_testing_bd_train	dim(bd_test)
400	200
11	11

	age	sex	region	income	married	children	car	save_act
76	fifties	FEMALE	RURAL	(2.44e+04,4.38e+04]	married=NO	2	car=YES	save_act=N
298	teens	MALE	RURAL	(5.01e+03,2.44e+04]	married=YES	2	car=NO	save_act=N
414	twenties	FEMALE	INNER_CITY	(5.01e+03,2.44e+04]	married=YES	0	car=YES	save_act=N
473	old	MALE	INNER_CITY	(5.01e+03,2.44e+04]	married=YES	2	car=YES	save_act=Y

	age	sex	region	income	married	children	car	save_act
325	thirties	MALE	SUBURBAN	(2.44e+04,4.38e+04]	married=YES	0	car=NO	save_act=YE
499	fifties	FEMALE	TOWN	(4.38e+04,6.31e+04]	married=NO	0	car=NO	save_act=YE

	age	sex	region	income	married	children	car	save_act
76	fifties	FEMALE	RURAL	(2.44e+04,4.38e+04]	married=NO	2	car=YES	save_act=NO
298	teens	MALE	RURAL	(5.01e+03,2.44e+04]	married=YES	2	car=NO	save_act=NO
414	twenties	FEMALE	INNER_CITY	(5.01e+03,2.44e+04]	married=YES	0	car=YES	save_act=NO
473	old	MALE	INNER_CITY	(5.01e+03,2.44e+04]	married=YES	2	car=YES	save_act=YE
325	thirties	MALE	SUBURBAN	(2.44e+04,4.38e+04]	married=YES	0	car=NO	save_act=YE
499	fifties	FEMALE	TOWN	(4.38e+04,6.31e+04]	married=NO	0	car=NO	save_act=YE

Model Data Prep

```
## [1] "Imbalance checking for pep target variable"
##
##  pep=NO pep=YES
##    0.57    0.43
## [1] "Imbalance is not a problem here as the no and yes values are nearl
y balanced"
## Rows: 200
## Columns: 11
## $ age         <int> 59, 18, 26, 61, 36, 51, 42, 46, 32, 55, 45, 23, 47,
37,...
## $ sex         <dbl> 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1,
1, 1...
## $ region      <fct> RURAL, RURAL, INNER_CITY, INNER_CITY, SUBURBAN, TOW
N, I...
## $ income      <dbl> 35611, 8639, 16519, 21140, 28495, 43800, 17390, 171
49, ...
```

```
## $ married     <dbl> 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1,
1, 0...
## $ children    <int> 2, 2, 0, 2, 0, 0, 0, 1, 0, 0, 0, 3, 2, 0, 1, 0, 0,
0, 0...
## $ car         <dbl> 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0,
1, 0...
## $ save_act    <dbl> 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,
1, 1...
## $ current_act <dbl> 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1...
## $ mortgage    <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0,
0, 1...
## $ pep         <fct> YES, NO, YES, NO, NO, NO, NO, YES, YES, YES, NO, NO
, YE...
```

age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
59	0	RURAL	35611	0	2	1	0	0	0	YES
18	1	RURAL	8639	1	2	0	0	0	0	NO
26	0	INNER_CITY	16519	1	0	1	0	1	0	YES
61	1	INNER_CITY	21140	1	2	1	1	0	0	NO
36	1	SUBURBAN	28495	1	0	0	1	1	0	NO
51	0	TOWN	43800	0	0	0	1	1	1	NO

```
## Rows: 200
## Columns: 12
## $ id          <fct> ID12176, ID12398, ID12514, ID12573, ID12425, ID1259
9, I...
## $ age         <int> 59, 18, 26, 61, 36, 51, 42, 46, 32, 55, 45, 23, 47,
37,...
## $ sex         <dbl> 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1,
1, 1...
## $ region      <fct> RURAL, RURAL, INNER_CITY, INNER_CITY, SUBURBAN, TOW
N, I...
## $ income      <dbl> 35611, 8639, 16519, 21140, 28495, 43800, 17390, 171
49, ...
```

```
## $ married     <dbl> 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1,
1, 0...
## $ children    <int> 2, 2, 0, 2, 0, 0, 0, 1, 0, 0, 0, 3, 2, 0, 1, 0, 0,
0, 0...
## $ car         <dbl> 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0,
1, 0...
## $ save_act    <dbl> 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1,
1, 1...
## $ current_act <dbl> 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1...
## $ mortgage    <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0,
0, 1...
## $ pep         <fct> YES, NO, YES, NO, NO, NO, NO, YES, YES, YES, NO, NO
, YE...
```

id	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
ID12176	59	0	RURAL	35611	0	2	1	0	0	0	YES
ID12398	18	1	RURAL	8639	1	2	0	0	0	0	NO
ID12514	26	0	INNER_CITY	16519	1	0	1	0	1	0	YES
ID12573	61	1	INNER_CITY	21140	1	2	1	1	0	0	NO
ID12425	36	1	SUBURBAN	28495	1	0	0	1	1	0	NO
ID12599	51	0	TOWN	43800	0	0	0	1	1	1	NO

Model Training

Apriori classifier was shown to be 100% accurate in predicting pep responders. This means we can use a classifier based upon aprior to determine the target audience if we assume pep response is a good index for propensity to respond.

CBA classifer accuracy is 1 or 100% accuracy"

```
## CBA Classifier Object
## Class:
## Default Class: NA
## Number of rules: 21
```

```
## Classification method: first
## Description: CBA algorithm (Liu et al., 1998)
## [1] "CBA classifer accuracy is  1  or 100% accuracy"
```

Training Other Classifer Models: Random Forest and Support Vector Machines

The results of the association rule based predictions were quite high, so other classifiaction method will be used to challenge or confirm.

Model Testing

Testing for random forest and support vector machine classifers. The most important features for the Random forest algorithm can be seen below here. **Income** is the most important by far followed by children.

```
## rf variable importance
##
##                  Overall
## income           100.000
## children          72.324
## age               36.619
## mortgage          26.776
## save_act          21.867
## married           20.302
## sex                4.896
## regionRURAL        1.949
## car                1.327
## current_act        0.691
## regionSUBURBAN     0.275
## regionTOWN         0.000
```

Results

Model Training Results

Definition of the Terms related to measuring classification models:

- Positive (P) : Observation is positive (for example: is an apple).
- Negative (N) : Observation is not positive (for example: is not an apple).
- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Accuracy: * Classification Rate or Accuracy is given by the relation: (TP + TN) / (TP + TN + FP + FN) * It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

Recall: Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN). Recall is given by the relation: TP / (TP + FN)

AUC-ROC curve: This is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. **Higher the AUC, better the model** is at binary predictions. ### Support Vector Machine and Random Forest training results

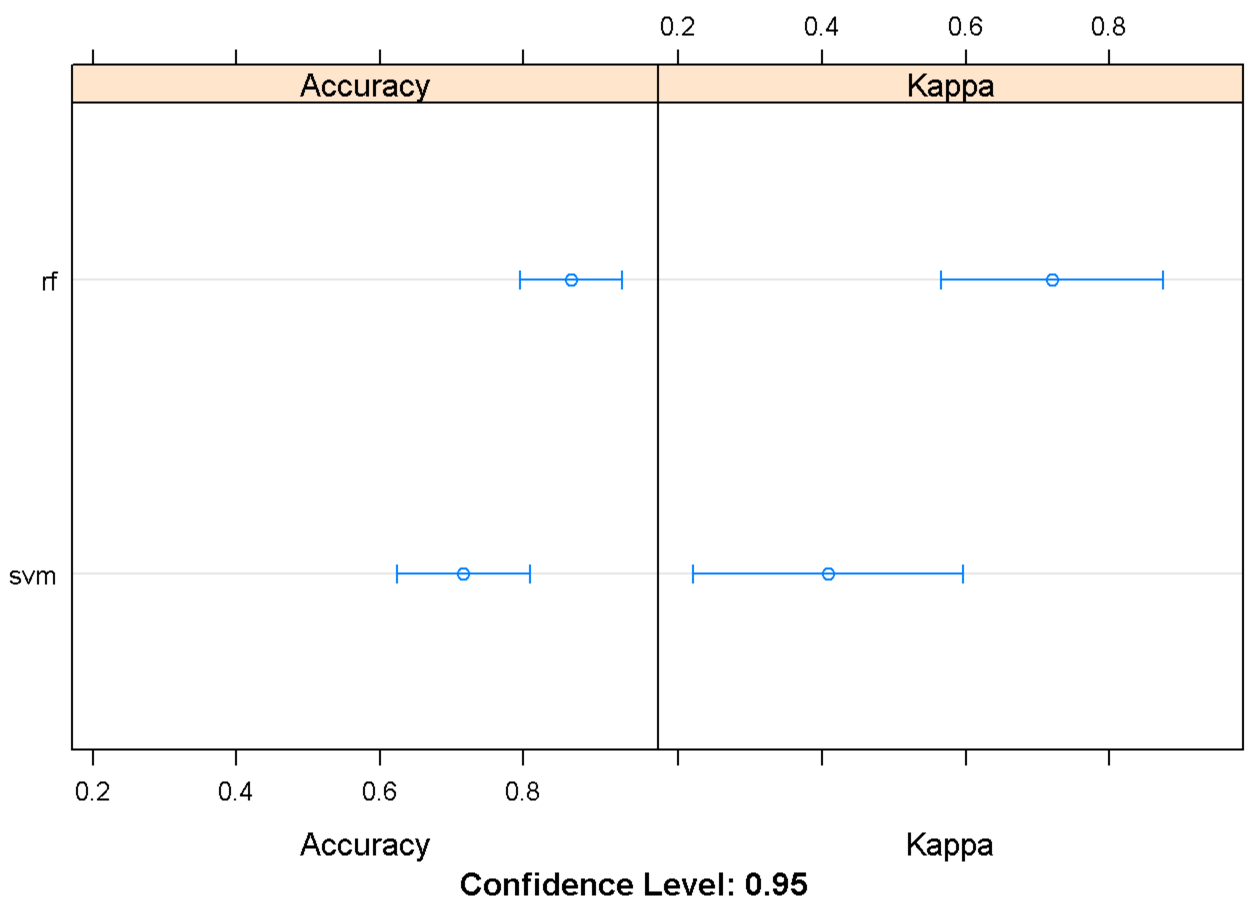
```
##
## Call:
## summary.resamples(object = results)
##
## Models: svm, rf
## Number of resamples: 10
##
## Accuracy
##     Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## svm 0.45    0.66   0.75 0.72    0.80 0.85    0
## rf  0.70    0.80   0.88 0.87    0.95 1.00    0
##
## Kappa
##      Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## svm -0.10    0.27   0.48 0.41    0.59 0.69    0
## rf   0.32    0.60   0.75 0.72    0.90 1.00    0
```

Support Vector Machine and Random Forest plotting results

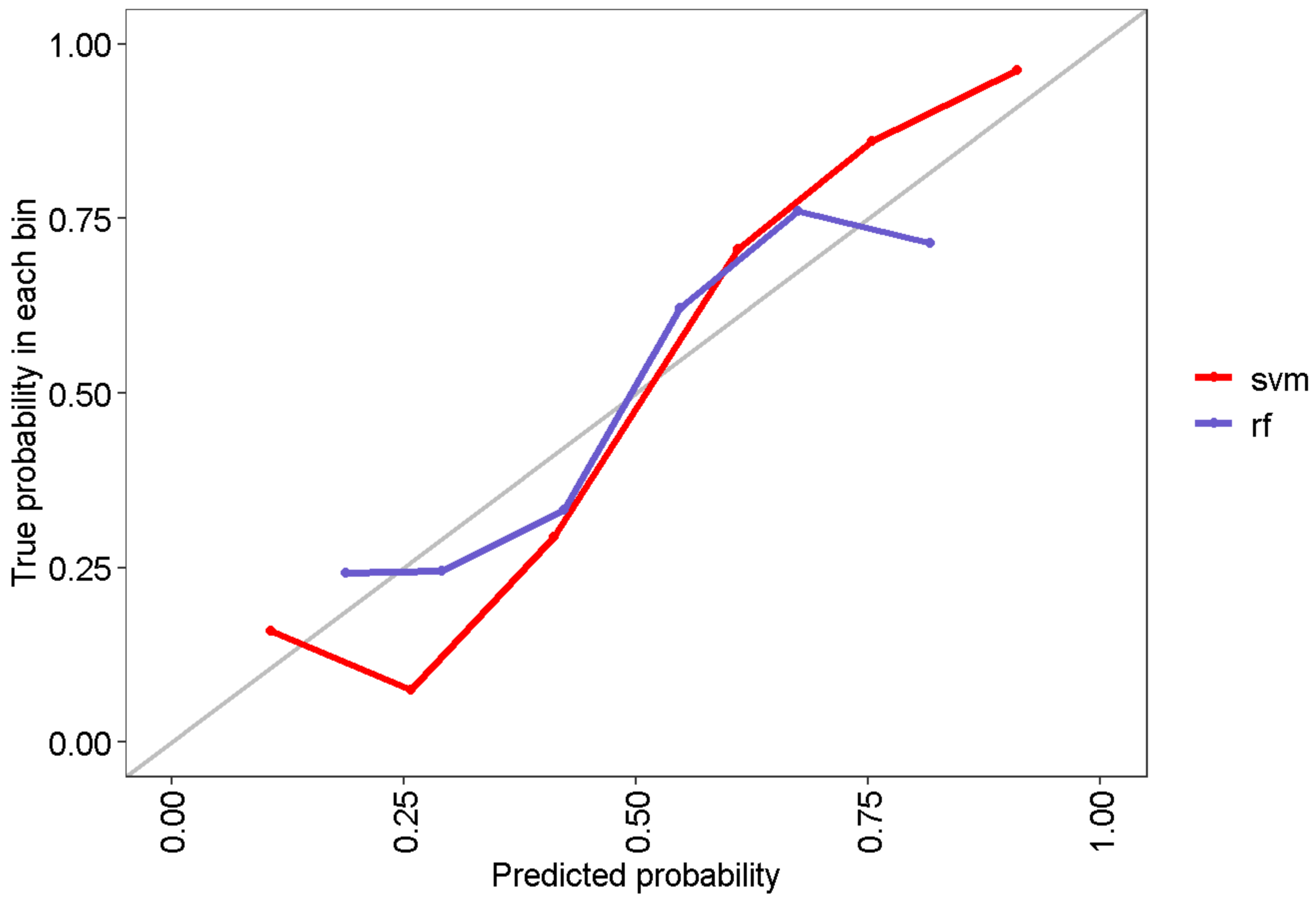
The graph below shows two evaluation metrics for the models.

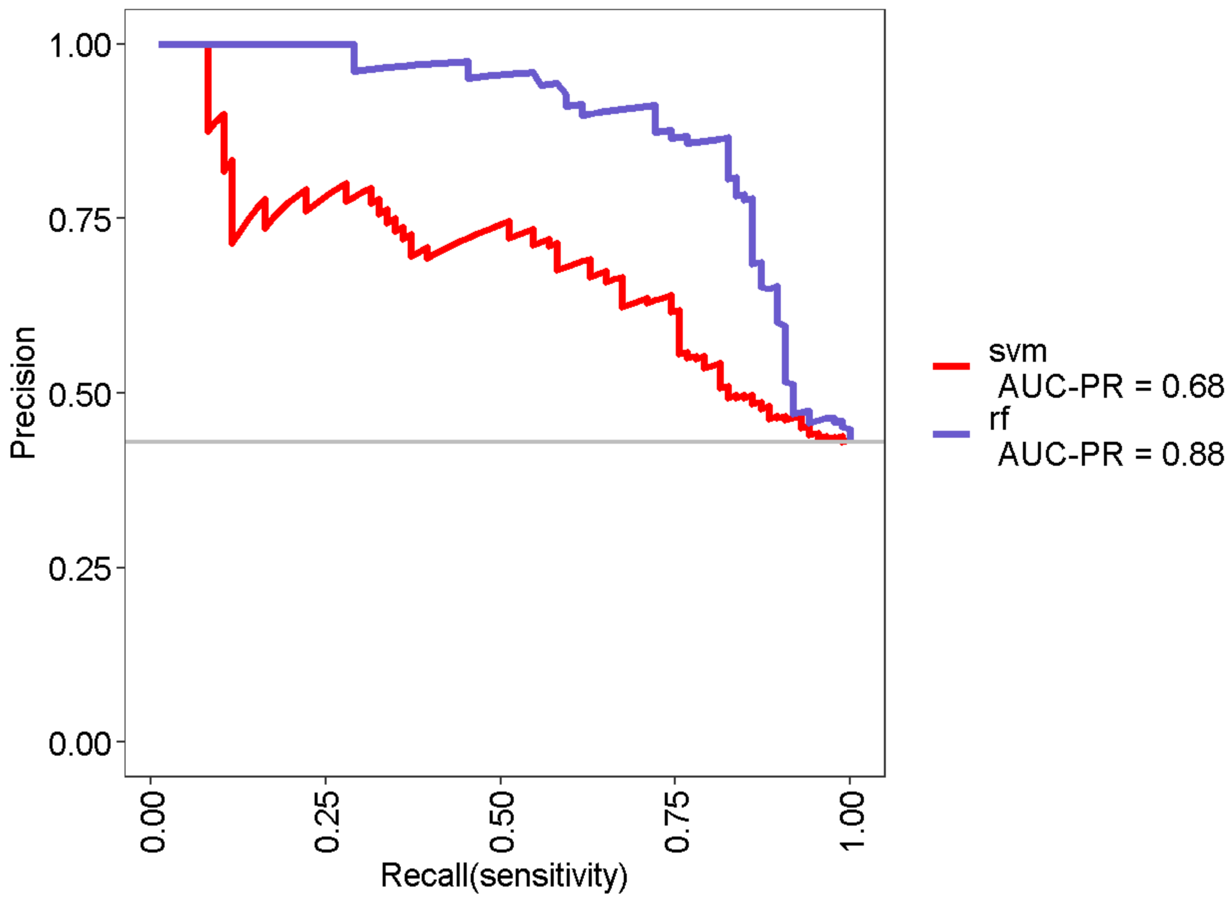
- Accuracy is the percentage of correctly classifies instances out of all instances. It is more useful on a binary classification than multi-class classification problems because it can be less clear exactly how the accuracy breaks down across those classes (e.g. you need to go deeper with a confusion matrix). Learn more about Accuracy here.
- Kappa or Cohen's Kappa is like classification accuracy, except that it is normalized at the baseline of random chance on your dataset. It is a more useful measure to use on problems that have an imbalance in the classes (e.g. 70-30 split for classes 0 and 1 and you can achieve 70% accuracy by predicting all instances are for class 0). Learn more about Kappa here.

The random forest algorithm denoted by *rf* is the better model according to these metrics.

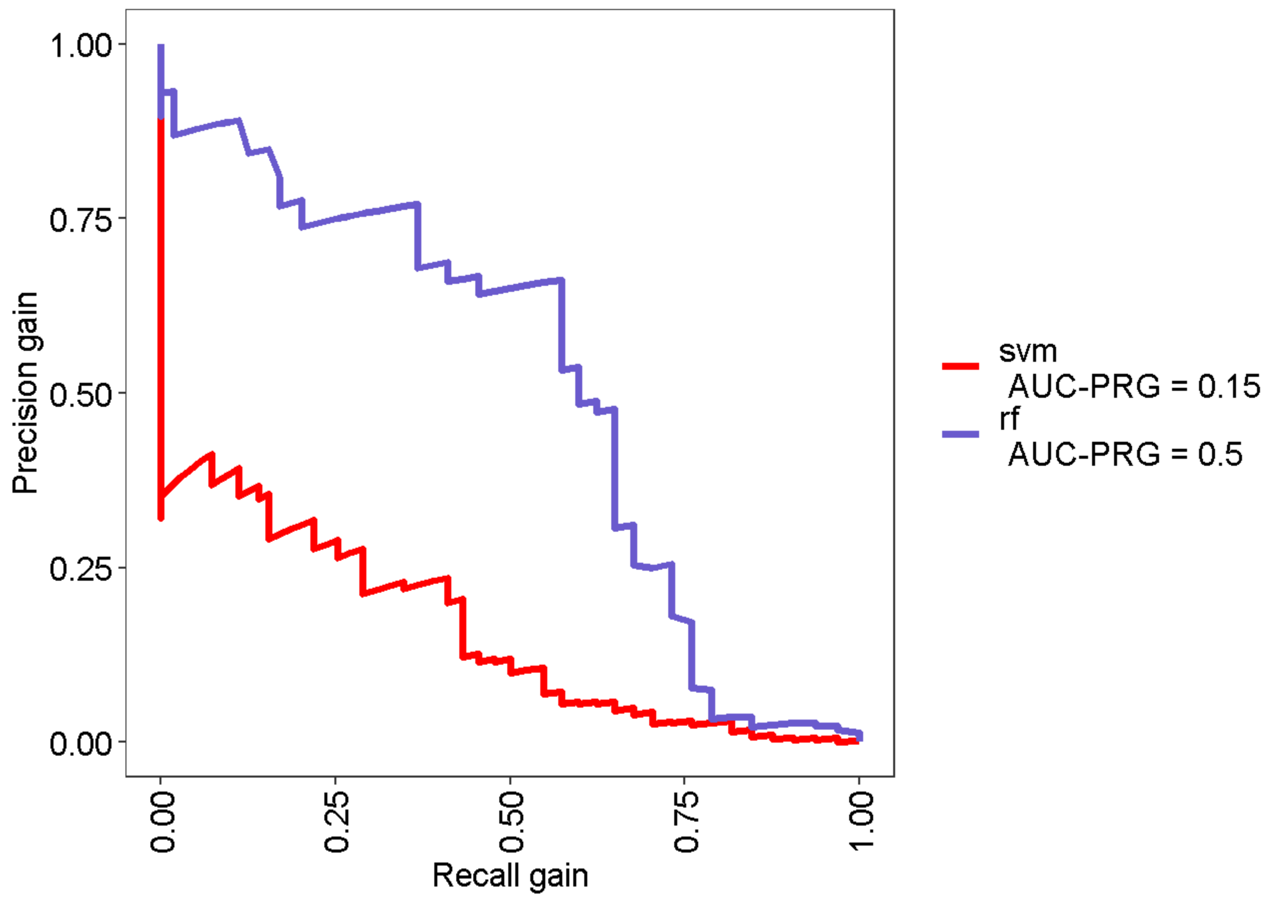


The last graph is plotting AUC-ROC and the random forest(rf) shows more area under the curve as compared to the support vector machine (svm) denoted by the red line. This means the random forest is a better fit for the data by this AUc-ROC metric.

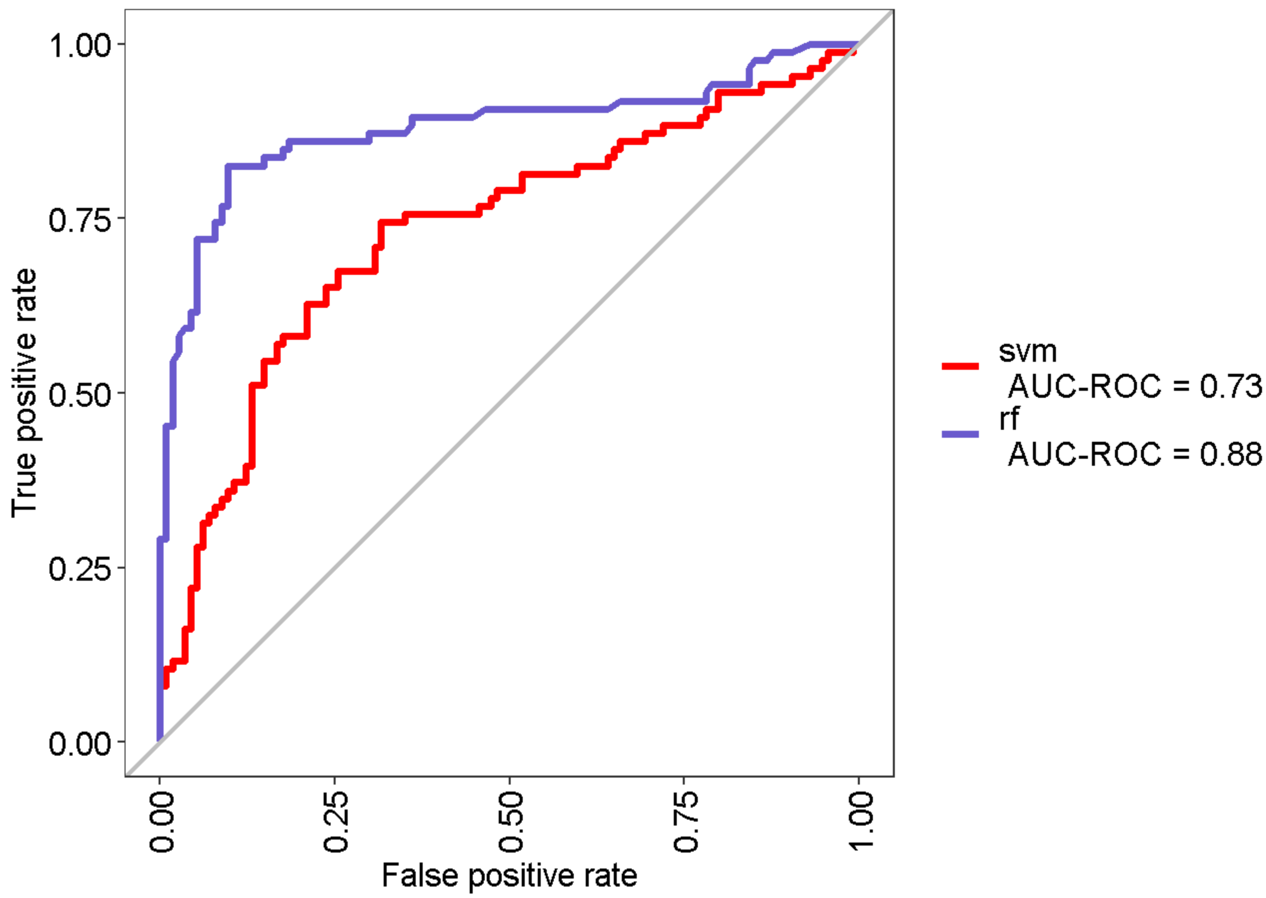

True probability in each bin
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Predicted probability
svm
rf



1.00
0.75
0.50
0.25
0.00
Precision
0.00
0.25
0.50
0.75
1.00
Recall(sensitivity)
svm
AUC-PR = 0.68
rf
AUC-PR = 0.88



1.00
0.75
0.50
0.25
0.00
Precision gain
0.00
0.25
0.50
0.75
1.00
Recall gain
svm
AUC-PRG = 0.15
rf
AUC-PRG = 0.5



Model Testing Results

Confusion Matrix

The random forest classifer was also able to get 100% accuracy predicting pep response using the test dataset.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  NO YES
##        NO  114   0
##        YES   0  86
##
##                Accuracy : 1
```

```
##                    95% CI : (0.982, 1)
##     No Information Rate : 0.57
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.00
##             Specificity : 1.00
##          Pos Pred Value : 1.00
##          Neg Pred Value : 1.00
##              Prevalence : 0.57
##          Detection Rate : 0.57
##    Detection Prevalence : 0.57
##       Balanced Accuracy : 1.00
##
##        'Positive' Class : NO
##
```

The support vector machine had significant as well with 85% accuracy.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  NO YES
##        NO  107  22
##        YES   7  64
##
##                Accuracy : 0.855
##                  95% CI : (0.798, 0.901)
##     No Information Rate : 0.57
##     P-Value [Acc > NIR] : < 2e-16
```

```
##
##                   Kappa : 0.698
##
##  Mcnemar's Test P-Value : 0.00933
##
##             Sensitivity : 0.939
##             Specificity : 0.744
##          Pos Pred Value : 0.829
##          Neg Pred Value : 0.901
##              Prevalence : 0.570
##          Detection Rate : 0.535
##    Detection Prevalence : 0.645
##       Balanced Accuracy : 0.841
##
##        'Positive' Class : NO
##
```

Conclusions

Association rule mining algorithms typically generate a large number of association rules, which poses a major problem for understanding and analyzing rules. The analysis performed here demonstrated the ability to to mine rules from the KNJ financial data using the apriori technique. Our process found 29 strong rules, but we filtered those rules down even further to our 5 strongest rules. The metrics measured to gauge rule strength were support, confidence and lift and there was an important filter used to consider past responders as the mostly likely to respond again.

The premise of past responders being the index for future responders was bolstered by using classification models. The classification models were trained to perdict passed responders by using the pep response feature. The classification approach simply tried to determine yes or no in terms of customer response After using several classification methods it is safe to say our premise of using past responses as a means to segment the customer, was an informed one. Two out of three models tested predicted with 100% accuracy, which isn't surprising given the strength of the rules we found using the apriori rule mining methods.

In summary, the KNJ Finance data supports a means to create rules and classifiaction models to build profiles, which can support a targeted marketing campaigns. The target customer segment is married couples with only 1 child who do business with KNJ, but not on the mortgage side. The top 5 rules were used to generate a targeted campaign with 29 accounts, and a broader campaign with 100 was created by using less rules. KNJ finanical will have the option to choose which campaign serves their purposes.