KENDRA OSBURN | 1-30-19 | IST707 | HW3 | ASSOCIATION RULE MINING .

PROJECT PURRRSONAL EQUITY PLAN

Furst Mutual Bank | Feline the Trend Data Analysis

NOTE FOR DR. GATES: *As you mentioned in your most recent lecture, practicing with meaningful data (or, as in this case, adding arbitrary meaning to the data) is helpful during the learning process. For the purposes of this assignment, I've decided that this is a world where cats are the dominant species (but almost everything else about the world -- besides some bad puns -- is the same).*

Introduction

Furst Mutual Bank, a relatively new bank on the block, had great success year one and is looking to get more of their clients invested in their Purrrsonal Equity Plans. Purrrsonal Equity Plans are beneficial both for the bank and the client, as this is an opportunity for the client to passively invest in something low risk and high reward. Furthermore, this is additional fiduciary catpital for Furst Mutual Bank to invest on a global scale. Why is Furst Mutual Bank pursuing Purrrsonal Equity Plans (PEPs)? Excellent question. PEPs are a persuasive alternative to traditional investment opportunities of decades past. Previously, many clients would squirrel away their extra income in their fancy cat towers or even behind the litter box (NOTE: FDM does NOT conde either of these behaviors). Money outside the bank isn't making money. Every dollar inside FDM is like a smaller cat working for, and earning money on behalf of, the client.

Furst Mutual Bank reached out to Feline the Trend Data Analysis Company to get a better handle on which customers they should reach out to first to get this new initiative up and out of the box. With such a large undertaking, where should they start? Which customers were most likely to be interested in a PEP? What indicators in the existing data suggest a customer might pursue a PEP? Using the initial data from year one, what rules or associations can be from the customers that have PEPs? With the information FMB already had on hand, Feline the Trend Data Analysis set out to answer these questions, and provide Furst Mutual Bank with a clear roadmap for their PEP sales team in the upcoming year.

NOTE FOR DR. GATES 2: *The more time I spent playing around with this data, the more I began to really understand what ARM was actually doing. I'm VERY TEMPTED to go back and redo everything (why was I SO SPECIFIC about 40-year-olds with one child when pretty much "have child" leads to a PEP!? Annnnd I went back to update it. I also have wanted to change "the approach" of this report at least 12 times. Also, I have SO. MANY. QUESTIONS. Should I define Association Rule Mining?! Yes, right?! What visualizations should I show!? The plots seem confusing and scattered (but they are SO FUN to interact with… maybe in this Feline Future everything is interactive?!) And now, again, I want to go back and edit everything because I just started playing with "pep=no" and learning from that. Remember at the beginning of this paragraph when I said I thought I understood what ARM was actually doing? Yeah, I'm only scratching the surface. There is never enough time. Why isn't there ever enough time?*


PEP?

Analysis and Models

NOTE: This is a technical breakdown of how the results were achieved. Unless attempting to replicate the findings, this section can be safely overlooked by the FMB team.

First, the libraries for Association Rule Mining (ARM) were added to the script. Then, the data was read in from the csv file provided by FMB and examined using the structure function. Initially, the data looked like FIGURE 1 below. Before any cleaning or processing, the data consisted of a unique id, age (numeric), sex (“male” or “female”), region (“inner city”, “rural”, “suburban”, “town”), income (numeric), married (“yes” or “no”), children (0-4), car (“yes” or “no”), saving account (“yes” or “no”), current account (“yes” or “no”), mortgage (“yes” or “no”), and PEP (“yes” or “no”).

FIGURE 1: ORIGINAL STRUCTURE

$ id	Factor w/ 600 levels
$ age	int 48 40 51 23 57 57 22 58 37 54 ...
$ sex	Factor w/ 2 levels "FEMALE","MALE"
$ region	Factor w/ 4 levels "INNER_CITY","RURAL",..
$ income	num 17546 30085 16575 20375 50576 ...
$ married	Factor w/ 2 levels "NO","YES"
$ children	int 1 3 0 3 0 2 0 0 2 2 ...
$ car	Factor w/ 2 levels "NO","YES"
$ save_act	Factor w/ 2 levels "NO","YES"
$ current_act	Factor w/ 2 levels "NO","YES"
$ mortgage	Factor w/ 2 levels "NO","YES"
$ pep	Factor w/ 2 levels "NO","YES"

FIGURE 2: ORIGINAL DATA

	id	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
1	ID12101	48	FEMALE	INNER_CITY	17546	NO	1	NO	NO	NO	NO	YES
2	ID12102	40	MALE	TOWN	30085	YES	3	YES	NO	YES	YES	NO
3	ID12103	51	FEMALE	INNER_CITY	16575	YES	0	YES	YES	YES	NO	NO
4	ID12104	23	FEMALE	TOWN	20375	YES	3	NO	NO	YES	NO	NO
5	ID12105	57	FEMALE	RURAL	50576	YES	0	NO	YES	NO	NO	NO
6	ID12106	57	FEMALE	TOWN	37870	YES	2	NO	YES	YES	NO	YES

DATA CLEANING

First, unnecessary data was removed. Not only was the ID unnecessary, it would disrupt our analysis. It was removed. NAs were also removed. Two categories, age and income were descritized. Age was made discrete by taking the continuous data and breaking it down into categories -- "child","teens","twenties","thirties","fourties","fifties","old". Income was made discrete by separating the continuous data into three categories "lowIncome", "midIncome", "highIncome". Finally, children were converted from numeric to nominal.

FIGURE 3: CLEANED STRUCTURE

```
'data.frame':   599 obs. of  11 variables:
 $ age        : Factor w/ 7 levels "child","teens",..: 5
 $ sex        : Factor w/ 2 levels "FEMALE","MALE": 1 2
 $ region     : Factor w/ 4 levels "INNER_CITY","RURAL",
 $ income     : Factor w/ 3 levels "lowIncome","midIncom
 $ married    : Factor w/ 2 levels "NO","YES": 1 2 2 2 2
 $ children   : Factor w/ 4 levels "0","1","2","3": 2 4
 $ car        : Factor w/ 2 levels "NO","YES": 1 2 2 1 1
 $ save_act   : Factor w/ 2 levels "NO","YES": 1 1 2 1 2
 $ current_act: Factor w/ 2 levels "NO","YES": 1 2 2 2 1
 $ mortgage   : Factor w/ 2 levels "NO","YES": 1 2 1 1 1
 $ pep        : Factor w/ 2 levels "NO","YES": 2 1 1 1 1
```

FIGURE 4: CLEANED DATA

	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
1	fourties	FEMALE	INNER_CITY	lowIncome	NO	1	NO	NO	NO	NO	YES
2	thirties	MALE	TOWN	midIncome	YES	3	YES	NO	YES	YES	NO
3	fifties	FEMALE	INNER_CITY	lowIncome	YES	0	YES	YES	YES	NO	NO
4	twenties	FEMALE	TOWN	lowIncome	YES	3	NO	NO	YES	NO	NO
5	fifties	FEMALE	RURAL	highIncome	YES	0	NO	YES	NO	NO	NO
6	fifties	FEMALE	TOWN	midIncome	YES	2	NO	YES	YES	NO	YES

ASSOCIATION RULE MINING

Association Rule Mining was used to create this analysis. Association Rule Mining (ARM) uses three key components to determine if something is salient enough to be considered “a rule” -- support, confidence and lift. Support is how frequently something appears within the dataset. Confidence is how frequently thing X appears in transactions that contain Y.

Lastly, there is lift. Simply put, lift is target response divided by average response. A high lift is something to pay attention to. A lift of one indicates independence. *NOTE FOR DR. GATES: I still don't REALLY understand what exactly is happening with lift no matter how many times I google it.*

ATTEMPTS

NOTE FOR DR. GATES: As I'm still so new to this whole process, I included all of my attempts at the "get to know you" part of the data analysis in hopes that if you see anything glaring that I could be doing better/differently, you'd point it out in the comments :)

ATTEMPT ONE

- bankRules = apriori(bankdata, parameter = list(supp = 0.001, conf = 0.9, maxlen = 3))
- options(digits=2)
- inspect(bankRules[1:40])
- rulesByLift <- head(sort(bankRules, by="lift"), 10)
- plot(rulesByLift, method="graph", interactive=TRUE)
- inspect(rulesByLift)

ATTEMPT TWO

- ## Changing confidence from 0.9 to 1
- bankRulesTwo = apriori(bankdata, parameter = list(supp = 0.001, conf = 1, maxlen = 3))
- options(digits=2)
- rulesByLiftTwo <- head(sort(bankRulesTwo, by="lift"), 10)
- inspect(rulesByLiftTwo)

ATTEMPT THREE

- ## Changing support > 0.001 to 0.01
- bankRulesThree = apriori(bankdata, parameter = list(supp = 0.01, conf = 1, maxlen = 3))
- options(digits=2)
- rulesByLiftThree <- head(sort(bankRulesThree, by="lift"), 10)
- inspect(rulesByLiftThree)

ATTEMPT FOUR

- bankRulesFour = apriori(bankdata, parameter = list(supp = 0.01, conf = 1, maxlen = 3))
- options(digits=2)
- rulesByLiftFour <- head(sort(bankRulesFour, by="lift"), 10)

- inspect(rulesByLiftFour)

ATTEMPT FIVE

- ## -- Changed support > 0.01 to 0.1
- bankRulesFive = apriori(bankdata, parameter = list(supp = 0.1, conf = 1, maxlen = 4))
- options(digits=2)
- rulesByLiftFive <- head(sort(bankRulesFive, by="lift"), 10)
- inspect(rulesByLiftFive)

```
> inspect(rulesByLiftFive)
    lhs                                  rhs             support confidence lift count
[1] {income=highIncome}                  => {save_act=YES} 0.13    1          1.5  80
[2] {income=highIncome,current_act=YES} => {save_act=YES} 0.11    1          1.5  63
```

ATTEMPT SIX

- ## -- Changed support > 0.1 to 0.05
- bankRulesSix = apriori(bankdata, parameter = list(supp = 0.05, conf = 1, maxlen = 3))
- options(digits=2)
- rulesByLiftSix <- head(sort(bankRulesSix, by="lift"), 10)
- inspect(rulesByLiftSix)

```
> inspect(rulesByLiftSix)
     lhs                                  rhs                support confidence lift count
[1]  {age=fourties,children=1}            => {pep=YES}          0.053   1          2.2  32
[2]  {age=teens}                          => {income=lowIncome} 0.062   1          2.1  37
[3]  {age=teens,current_act=YES}          => {income=lowIncome} 0.055   1          2.1  33
[4]  {income=highIncome}                  => {save_act=YES}     0.134   1          1.5  80
[5]  {age=old,income=highIncome}          => {save_act=YES}     0.082   1          1.5  49
[6]  {income=highIncome,children=0}       => {save_act=YES}     0.055   1          1.5  33
[7]  {region=INNER_CITY,income=highIncome} => {save_act=YES}    0.055   1          1.5  33
[8]  {income=highIncome,pep=YES}          => {save_act=YES}     0.090   1          1.5  54
[9]  {income=highIncome,car=YES}          => {save_act=YES}     0.073   1          1.5  44
[10] {sex=MALE,income=highIncome}         => {save_act=YES}     0.065   1          1.5  39
```

ATTEMPT SEVEN

- ## -- Changed support > 0.1 to 0.05
- bankRulesSeven = apriori(bankdata, parameter = list(supp = 0.1, conf = 0.9, maxlen = 3))
- options(digits=2)
- rulesByLiftSeven <- head(sort(bankRulesSeven, by="lift"), 10)
- inspect(rulesByLiftSeven)

```
> inspect(rulesByLiftSeven)
     lhs                                    rhs                 support confidence lift count
[1]  {age=twenties,current_act=YES}      => {income=lowIncome}  0.15    0.96       2.0   87
[2]  {age=twenties,car=NO}               => {income=lowIncome}  0.11    0.96       2.0   65
[3]  {age=twenties,region=INNER_CITY}    => {income=lowIncome}  0.10    0.95       2.0   62
[4]  {age=twenties}                      => {income=lowIncome}  0.19    0.95       2.0  112
[5]  {age=twenties,married=YES}          => {income=lowIncome}  0.12    0.95       2.0   73
[6]  {age=twenties,save_act=YES}         => {income=lowIncome}  0.11    0.94       2.0   68
[7]  {age=twenties,mortgage=NO}          => {income=lowIncome}  0.13    0.94       2.0   76
[8]  {age=twenties,pep=NO}               => {income=lowIncome}  0.12    0.94       2.0   73
[9]  {income=highIncome}                 => {save_act=YES}      0.13    1.00       1.5   80
[10] {income=highIncome,current_act=YES} => {save_act=YES}      0.11    1.00       1.5   63
```

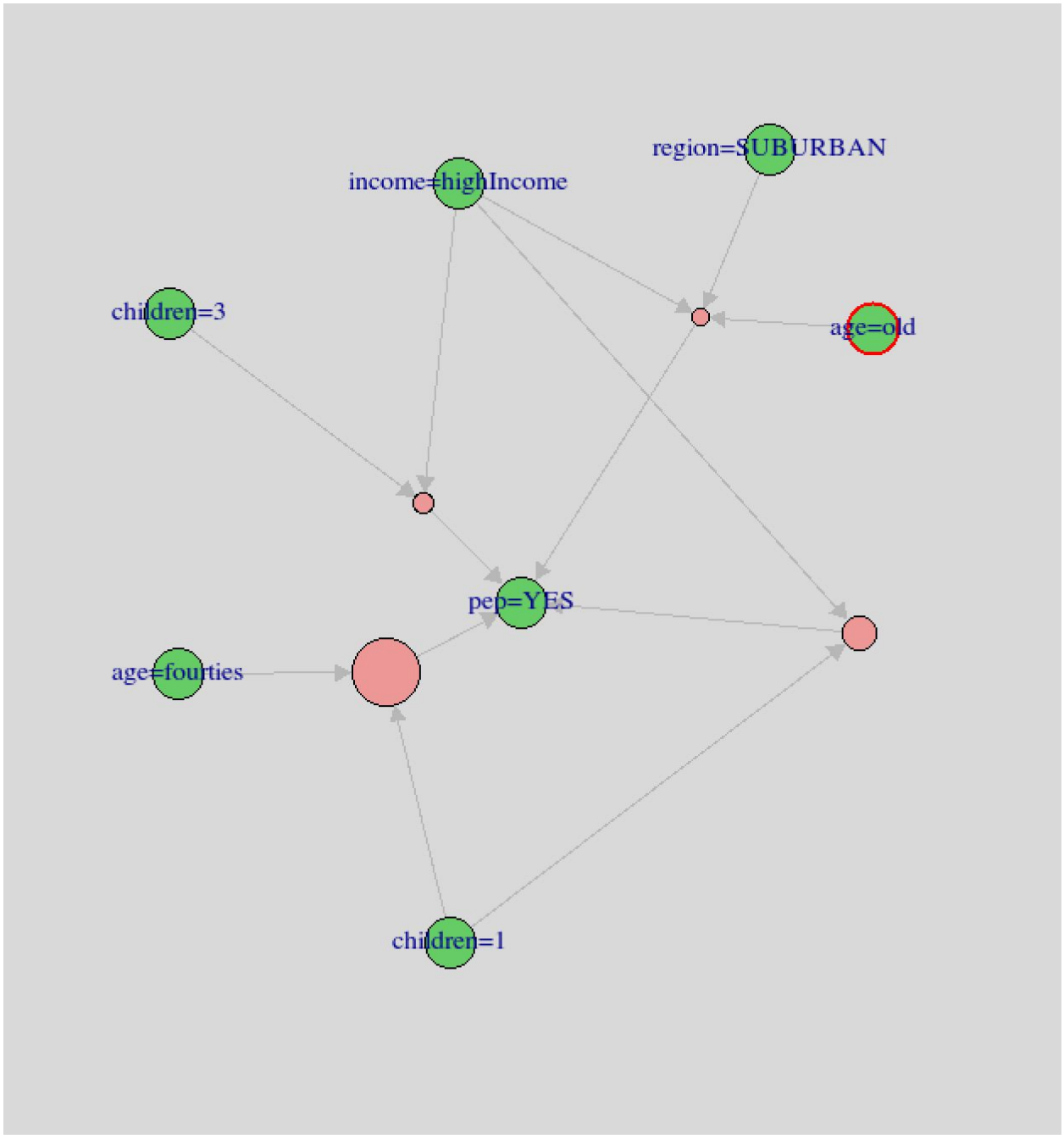
ATTEMPT EIGHT

- ## -- Changed CONFIDENCE > 0.9 to 0.2
- bankRulesEight = apriori(bankdata, parameter = list(supp = 0.01, conf = 0.2, maxlen = 3))
- options(digits=2)
- rulesByLiftEight <- head(sort(bankRulesEight, by="lift"), 10)
- inspect(rulesByLiftEight)

```
> inspect(rulesByLiftEight)
     lhs                                    rhs                 support confidence lift count
[1]  {income=highIncome,children=3}      => {region=RURAL}      0.013   1.00       6.2    8
[2]  {region=TOWN,income=highIncome}     => {age=old}           0.022   0.81       5.4   13
[3]  {age=old,children=2}                => {income=highIncome} 0.030   0.72       5.4   18
[4]  {income=highIncome,children=2}      => {age=old}           0.030   0.78       5.2   18
[5]  {age=old,pep=YES}                   => {income=highIncome} 0.062   0.69       5.1   37
[6]  {income=highIncome,children=1}      => {age=old}           0.020   0.75       5.0   12
[7]  {age=old,region=SUBURBAN}           => {income=highIncome} 0.012   0.64       4.8    7
[8]  {age=old,save_act=YES}              => {income=highIncome} 0.082   0.64       4.8   49
[9]  {children=3,pep=YES}                => {income=highIncome} 0.013   0.62       4.6    8
[10] {income=highIncome,pep=YES}         => {age=old}           0.062   0.69       4.6   37
```

(there were many more attempts)

ATTEMPTS WITH PEP:


region=SUBURBAN
income=highIncome
children=3
age=old
pep=YES
age=fourties
children=1

```
> inspect(rulesRightPepByLift)
     lhs                                                rhs       support confidence lift count
[1]  {income=highIncome,children=3}                  => {pep=YES} 0.0134  1          2.2   8
[2]  {income=highIncome,children=1}                  => {pep=YES} 0.0267  1          2.2  16
[3]  {age=fourties,children=1}                       => {pep=YES} 0.0534  1          2.2  32
[4]  {age=teens,region=SUBURBAN,children=2}          => {pep=YES} 0.0017  1          2.2   1
[5]  {age=teens,region=SUBURBAN,mortgage=YES}        => {pep=YES} 0.0017  1          2.2   1
[6]  {age=teens,region=SUBURBAN,children=0}          => {pep=YES} 0.0017  1          2.2   1
[7]  {age=teens,region=RURAL,children=1}             => {pep=YES} 0.0017  1          2.2   1
[8]  {age=teens,region=RURAL,mortgage=YES}           => {pep=YES} 0.0017  1          2.2   1
[9]  {age=teens,current_act=NO,mortgage=YES}         => {pep=YES} 0.0017  1          2.2   1
```

Analysis can sometimes be misleading. Take for example this output. It might be tempting to conclude that young teen parents are a good target, however, it is important to remember to take everything, including count, into consideration.

Results

WTF does that all mean!?

Five Rules

1. **Clients with one child have 81% chance of also having a PEP.**
    - SUPPORT: 0.18 | CONFIDENCE: 0.81 | LIFT: 1.8
2. **Clients who earn a high income with 1+ children have 96%+ chance of also having a PEP.**
    -
```
[2] {income=highIncome,children=3}      => {pep=YES} 0.013   1.00       2.2   8
[3] {income=highIncome,children=2}      => {pep=YES} 0.037   0.96       2.1  22
[4] {income=highIncome,children=1}      => {pep=YES} 0.027   1.00       2.2  16
```
3. **Clients who are in their 40s with one child, who also have either a savings account or a checking account, have <u>very high</u> chance of also having a PEP**
    -
```
[5] {age=fourties,married=YES,children=1}      => {pep=YES} 0.032   1          2.2  19
[6] {age=fourties,children=1,save_act=YES}     => {pep=YES} 0.043   1          2.2  26
[7] {age=fourties,children=1,current_act=YES}  => {pep=YES} 0.042   1          2.2  25
```
4. **Clients who earn a high income and live in a suburban area are <u>very</u> likely to have a PEP**
    -
```
[4]  {age=old,region=SUBURBAN,income=highIncome}      => {pep=YES} 0.012   1.00       2.2   7
[5]  {region=SUBURBAN,income=highIncome,children=2}   => {pep=YES} 0.010   1.00       2.2   6
[6]  {sex=MALE,region=SUBURBAN,income=highIncome}     => {pep=YES} 0.012   1.00       2.2   7
```
5. **Teenagers are 70% likely to NOT have a PEP**
    - [“pep=no”] SUPPORT: 0.043 | CONFIDENCE: 0.70 | LIFT | 1.3

NOTE TO DR. GATES: There are so many more ways I want to explore and play with this data. I feel like I'm just getting started (20+ hours in). I want to explore putting different things in the lhs and sort by those things, GROUP by other things (like CHILD and NO CHILD), use different bins etc… PRUNE some other things… I couldn't figure out how to prune things so that ABC->D and AB->D weren't both showing up :(

Conclusion

TOP LEVEL FINDINGS -- WHO *(AND WHO NOT TO)* TARGET

WHO TO TARGET:

1. GENERAL: **Target potential customers who have at least one child.** *Anyone with a child is 81% likely to have a PEP.*
2. MORE FOCUSED: **Target high income earners who have children.** *Very high lift meaning frequently in the dataset with a very strong correlation.*
3. EVEN MORE FOCUSED: **Target 40-year-olds with 1 child who already have a savings account or checking account.** *Very high lift meaning frequently in the dataset with a very strong correlation.*

WHERE TO TARGET:

4. GEOGRAPHIC: **Target those living in suburban areas.** *Across all models, this bubbled to the top.*

NO NEED TO TARGET:

5. NOT TO TARGET: **Teenagers.** *70% of teenagers are likely to not have a PEP account. Note: In the human world, only those 18+ are eligible for PEP accounts but in the feline world, who knows. It was simply valuable to turn the rhs from "pep=yes" to "pep=no"*

With PEP on the right and sorting by Lift, trends like "high income," "with children," "suburban" and "forties" continued to show up. This isn't entirely surprising and this information should be used to affirm Furst Mutual Banks current marketing strategy of targeting the upper-middle-class suburbanites.

As a full-service analytics company, Feline the Trend Data Analytics traditionally leaves their clients with a skeletal roadmap for the upcoming months. This is not a requirement as

much as it is a place for internal discussions to begin. Feline the Trend recommends focussing on the “in-house-wins” for Q1. This means identifying the 40-year-old parents within the system who currently have either a savings account or checking account, but no PEP, and sending a postcard with Feline the Trend Data Analytics proprietary CatnipCardboard ™ with information on how and why they should invest in a PEP. Feline the Trend recommends sending CatnipCardboard ™ once every quarter to this cohort. In Q2, Feline the Trend recommends taking a slightly more aggressive and inventive approach to entice cohort 2 -- the high income parents. Feline the Trend suggests identifying local private schools and offering to sponsor some of the sporting events with Feline the Trend Porta-Pissers ™. These portable litter boxes are an excellent place to showcase the benefits of PEP to a captive audience. In Q3 Feline the Trend recommends hosting a Signature CatnipCooler ™ event at a local watering hole to captivate some of the suburbanites who might have missed the first two memos. For further information about any of these recommendations, please contact Cindy at cindy@felinethetrend.com or 4242 Cat Hair Way.

Furst Mutual Bank has an exciting year ahead. As this report demonstrates, there are many different growth opportunities for Furst Mutual Bank in the upcoming months. Feline the Trend Data Analysis looks forward to an ongoing partnership.