Probabilistic Graphical Model (Bayesian Network)

Advanced Distributed Machine Learning Labs (ADML)

Essex Lake Group LLC

Knowledge Sharing Session

July, 2015

Objective: Embracing Bayesian approaches for business problem solving

Technical Objectives:

- How can we identify the probabilistic relationships among the variables in the given data
- For any given target, how can we identify the most influencing variables ?
- Can Bayesian networks help us in finding better insights than the existing Pivot approach ?

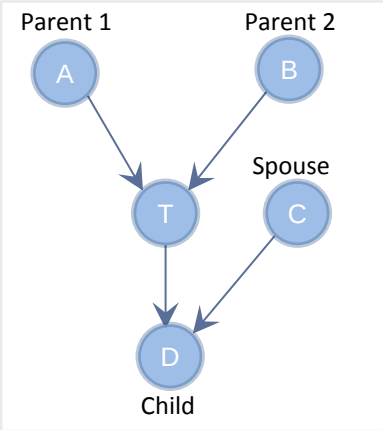
Business Objective :

- For the MET004 Account level data, how do we evaluate the performance of "*Participation Rate"* using PGM

Agenda

- **How do we construct Probabilistic Graphical Models (PGMs) ?**
- PGMs on Sales Ratio Data
- PGM with respect to CART

What are Probabilistic Graphical Models /Bayesian Networks?

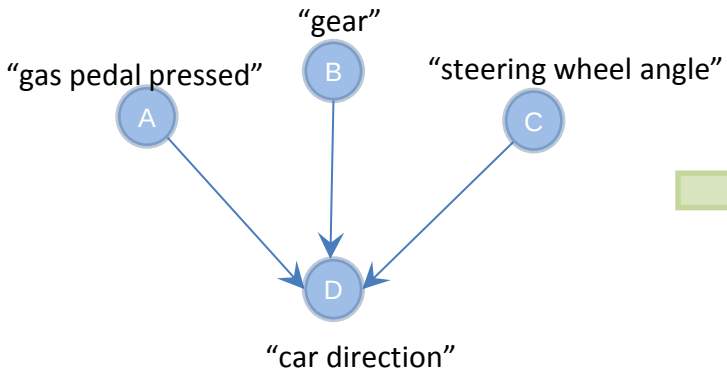


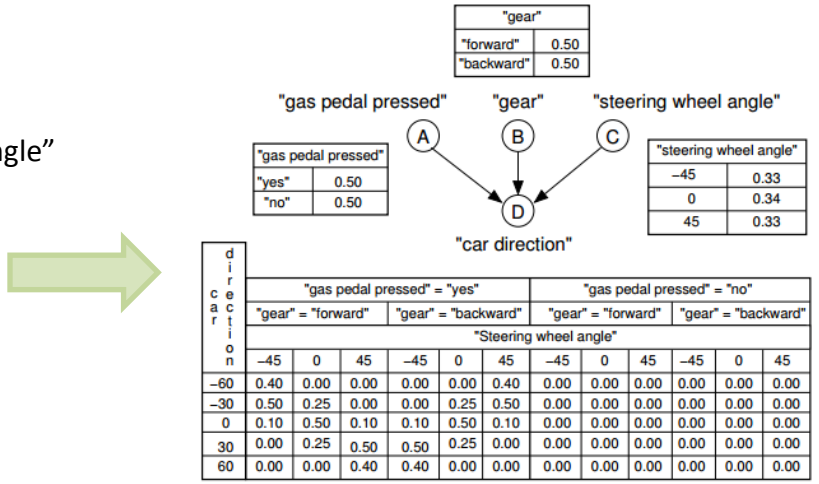
- Bayesian Network is Directed Acyclic Graph
- Nodes are variables and edges are relationships between variables.
- Each node has set of values
- Direction of edges indicates Causality i.e. cause-effect relationships among variables

Graph's Joint Probability Table will be tabulated using the following equation:

P(A, B, C, D) = $P_B(X_1, \ldots, X_n) = \prod_{i=1}^{n} P_B(X_i|\Pi_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\Pi_{X_i}}$

Example: "Car"





"gear"	
"forward"	0.50
"backward"	0.50

"gas pedal pressed"	
"yes"	0.50
"no"	0.50

"steering wheel angle"	
–45	0.33
0	0.34
45	0.33

car direction	"gas pedal pressed" = "yes"						"gas pedal pressed" = "no"					
	"gear" = "forward"			"gear" = "backward"			"gear" = "forward"			"gear" = "backward"		
	"Steering wheel angle"											
	–45	0	45	–45	0	45	–45	0	45	–45	0	45
–60	0.40	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.00
–30	0.50	0.25	0.00	0.00	0.25	0.50	0.00	0.00	0.00	0.00	0.00	0.00
0	0.10	0.50	0.10	0.10	0.50	0.10	0.00	0.00	0.00	0.00	0.00	0.00
30	0.00	0.25	0.50	0.50	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
60	0.00	0.00	0.40	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

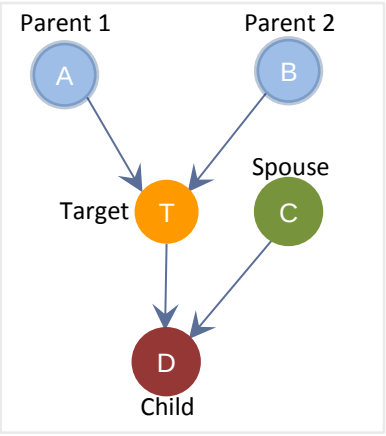
- They are **graphical** models, capable of displaying relationships clearly and intuitively
- They are **directional**, thus being capable of representing cause-effect relationships
- They **handle uncertainty** though the established theory of probability
- They can be used to represent **indirect** in addition to **direct causation**

Network Terminology

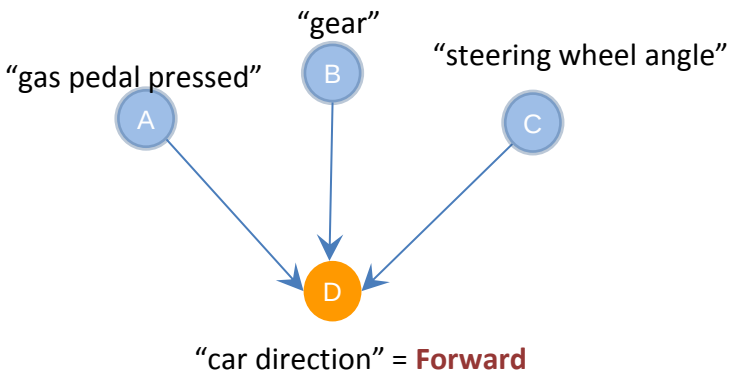
Markov Blanket: For a given target variable T **What are the most influencing variables ?**

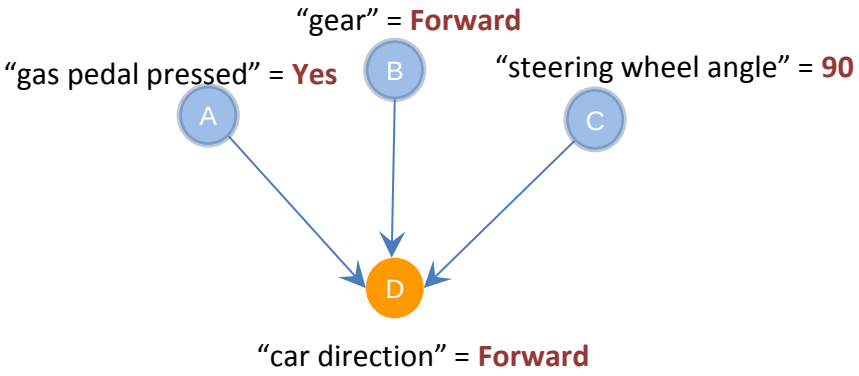
The **Markov Blanket** of a (Target) node consists of all nodes that make this Target conditionally independent of all the other nodes in the model:

- The ***Parents*** (blue nodes in the example) are used for cutting the information coming from their ascendants
- The ***Children*** (red node) are used for cutting the information coming from their descendants
- The ***Spouses*** (or co-parents, dark green) are used for cutting the information coming from the ascendants of the Children



Dependency among the variables of a system (Car)





If one Variable takes on some value then other variables are going to take that value

Constructing Bayesian Network - Process Flow

01. Processing data

- Data obtained after processing must have:
  - categorical values
  - must be binned in case of continuous values
- All data must be converted to factors.

02. Creating relationships between variables

- For the fitting of Bayesian Network, a scoring function ϕ is considered
- Given data and scoring function ϕ, find a network such that it maximizes the value of ϕ.
- Model tries various permutations of adding and removing arcs, such that final network has maximum posterior probability distribution, starting from a prior probability distribution on the possible networks, conditioned on data

03. Build CPTs of final fitted network

- While learning the network, the joint probability distribution is created for all the variables in the data set
- This joint probability distribution is converted to conditional probabilities for each dependent variable.
- These CPTs are used to solve any kind of analysis problems.

04. Record Network

- Import the learned structure and its CPT's into SAMIAM to form hypotheses and verify them.

How do we interpret the Bayesian Network for problem solving ?

Example: 1

A. Forward Analysis :

Q1. I have not done any mailing activity this year. How did this affect my participation percentage?

Q2. To check for conditional dependencies like mailing activity not being helpful if the account is already active.

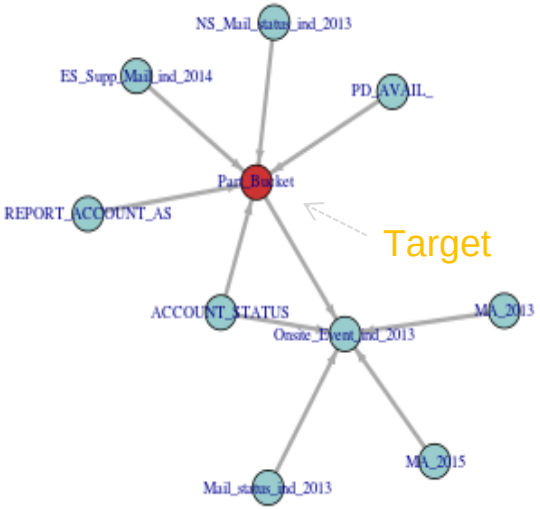
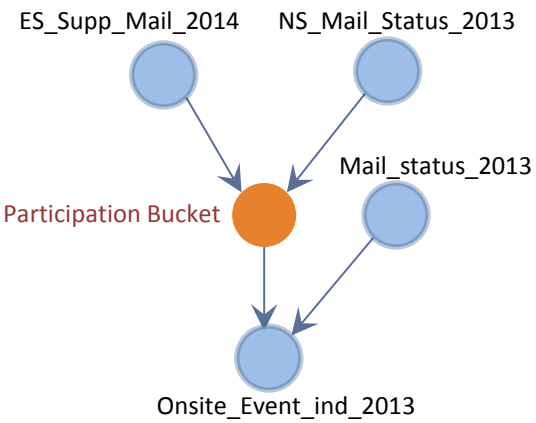


Fig: Markov Blanket for a target variable

B. Backward Analysis

Q3. If we know that our participation is high, then find responsible variables and their values to replicate the result in the future.

Q4. For my participation bucket to lie in the higher bucket, should I offer payroll deduct or not?



Variable Name	Evidence Set	Participation Bucket value before evidence	Participation Bucket value after evidence	Change in Values
Sponsored Mail 2014	Y	15.15%	19.46%	+4.31%
Non Sponsored Mail 2013	Y	15.15%	18.85%	+3.70%
Mail Status 2013	Y	15.15%	18.68%	+3.53%
Sponsored Mail 2014	N	15.15%	13.8%	-1.35%
Non Sponsored Mail 2013	N	15.15%	10.30%	-4.85%
Mail Status 2013	N	15.15%	10.32%	-4.83%

All Probability Values for Participation being in the 2.5%-5% (high) bucket.

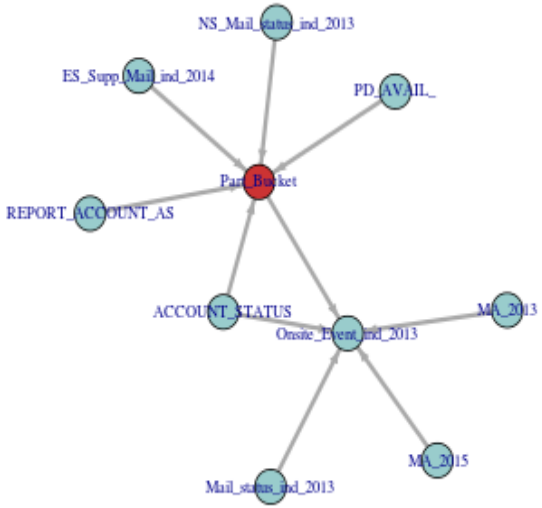
One more example → How do we interpret the Bayesian Network for problem solving ?

Example: 2

A. Forward Analysis :

Q1. I have not done any mailing activity this year. How did this affect my participation percentage?

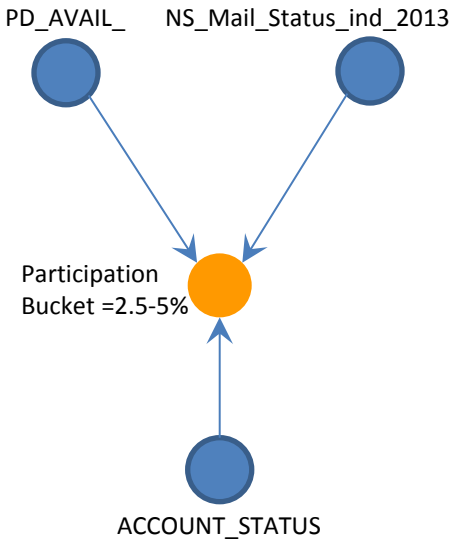
Q2. To check for conditional dependencies like mailing activity not being helpful if the account is already active.



B. Backward Analysis

Q3. If we know that our participation is high, then find responsible variables and their values to replicate the result in the future.

Q4. For my participation bucket to lie in the higher bucket, should I offer payroll deduct or not?



Evidence Value	Non-Sponsored Mail 2013	Payroll Deduct	Account Status
Participation Bucket = 2.5-5%	Y	Y	ACTIVE

Participation Bucket	Variable	Evidence	Participation Bucket Value before Evidence	Participation Bucket Value before Evidence	Change in Values
2.5% - 5%	Non-Sponsored Mail 2013 Payroll Deduct Account Status	Y Y ACTIVE	15.15%	28.55%	+13.40%

All Probability Values for Participation being in the 2.5%-5% (high) bucket.

Most Probable Explanation for given variable (MAP): *Process Flow*

01. Selection of Target Variable

- The network consists of all the variables in the data set
- To perform any kind of analysis, the target variable must be selected.
- The network is built without keeping a target variable so in case analysis needs to be done on any other variable, it can also be selected in the same network.

02. Find Target's Markov Blanket

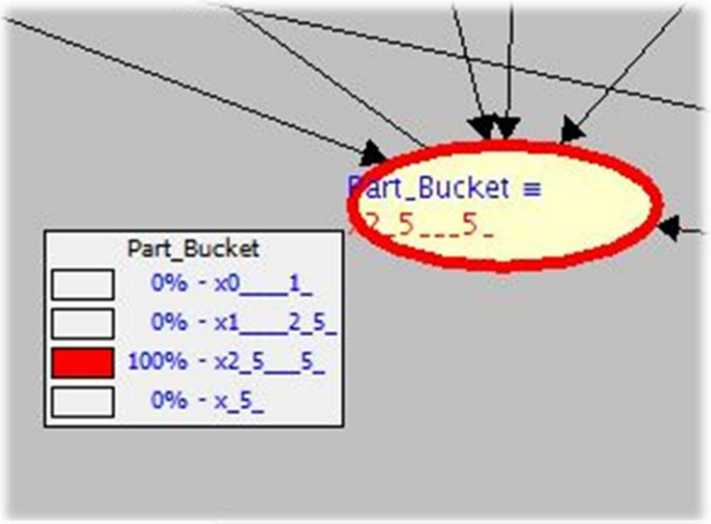
- Target's parent nodes, child nodes and all other parents of target's child nodes form it's Markov Blanket.
- These variables make the target variable independent from the rest of the variables in the network.

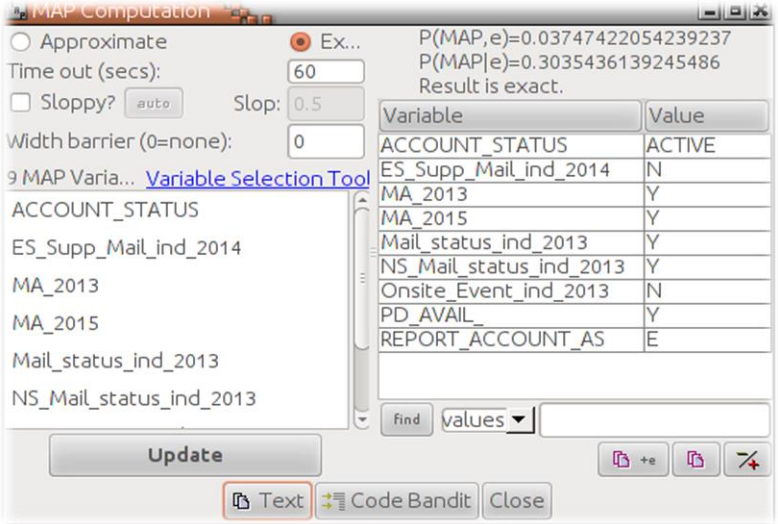
03. Set Evidence on Target

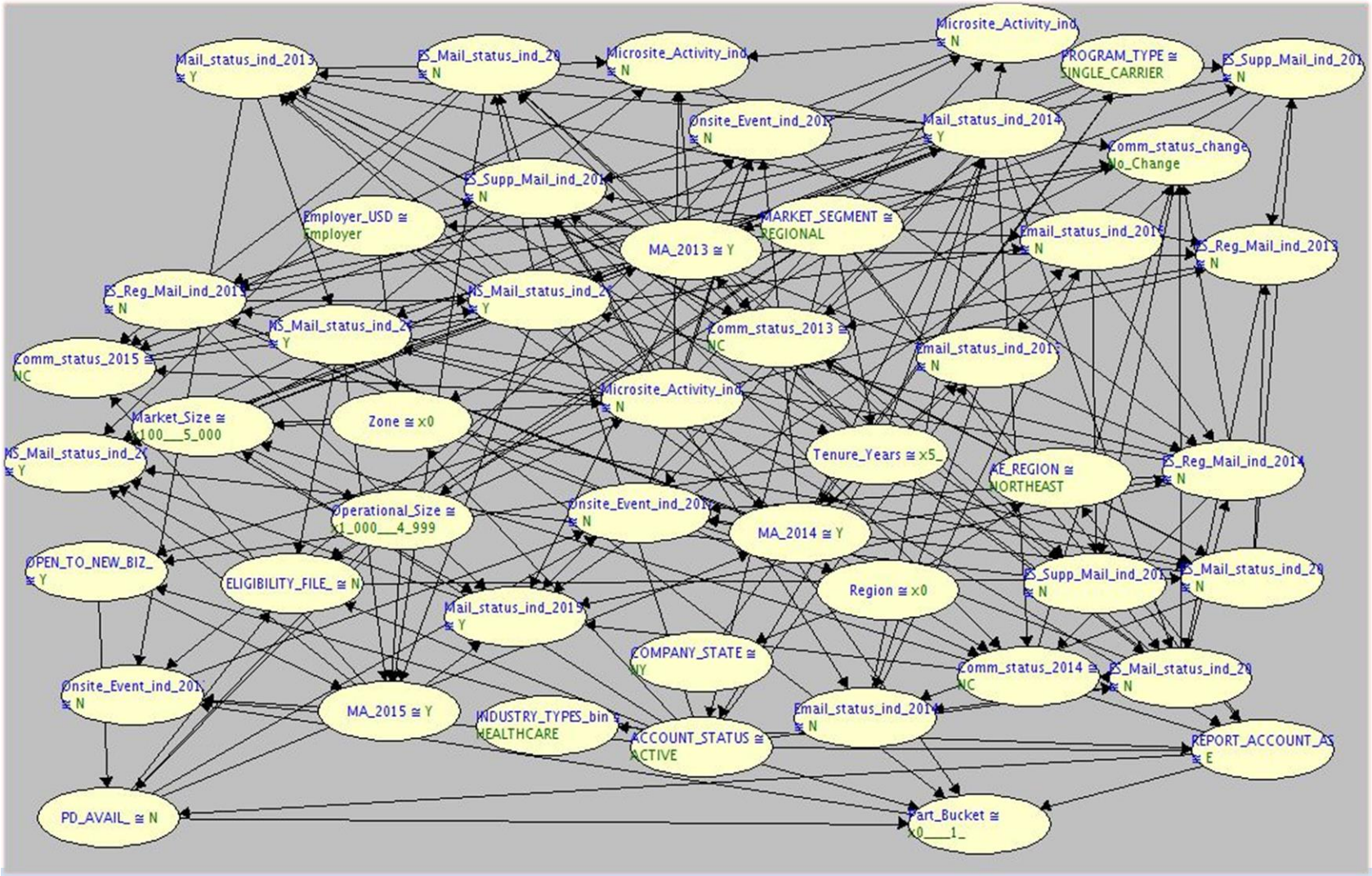
- In Samiam, for the given Bayesian Network, set evidence to the target variable, which is the variable of our interest.
- We need to find what values should other variables have so as to observe the evidenced value of target variable.

04. Run MAP Query on Markov Blanket

- Select the target variable's Markov Blanket variables for which MAP values are to be computed.
- Run MAP query.
- This query returns the value of the selected variables such that the posterior probability of the evidenced target is maximized.

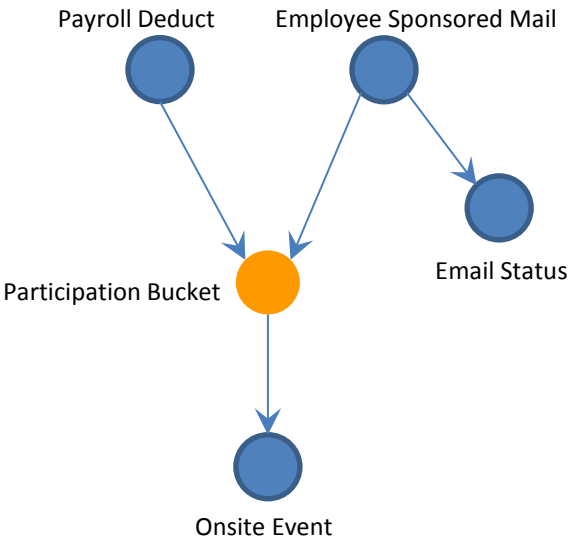






SAMIAM and R-SHINY Demo

More details on Network Structure:



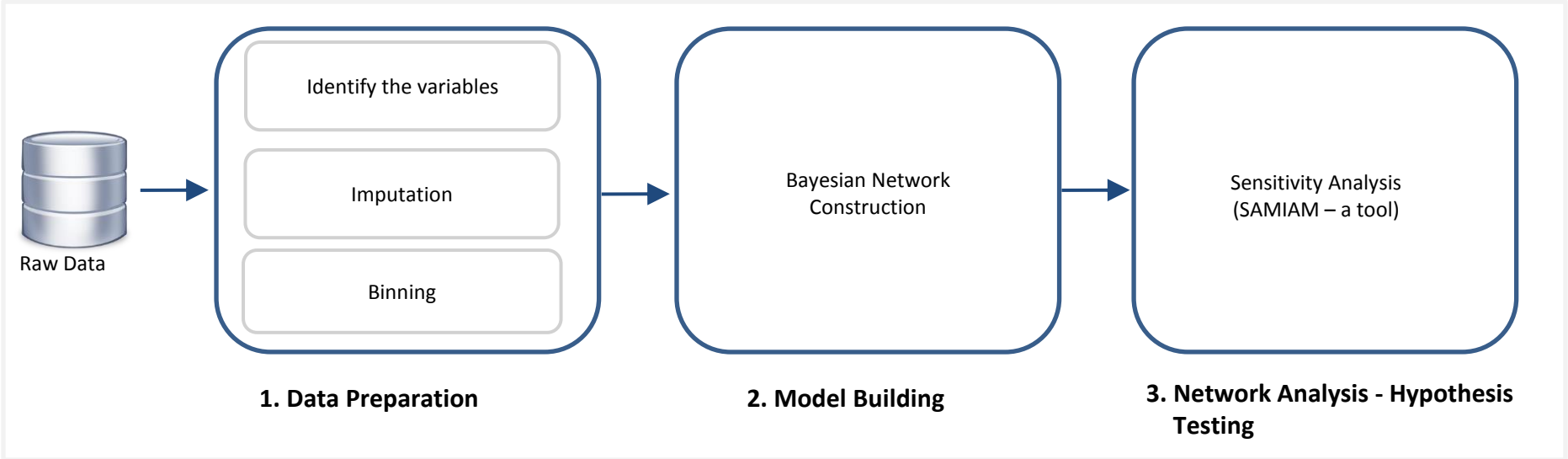
- Payroll Deduct, Employee Sponsored Mail and Participation Bucket for a V-Structure.
- This V-Structure means that Payroll Deduct and Employee Sponsored Mail are **independent** of each other.
- This means, any change in Payroll Deduct **won't influence** Employee Sponsored Mail and vice versa.
- Also, any change in Email Status which is the sibling of Participation Bucket will result in change in Participation Bucket as well due to the presence of an active trail between the two.
- Payroll deduct can however, **affect** Onsite Event through Participation Bucket but not Email Status due to the presence of a V-Structure between Payroll Deduct and Employee Sponsored Mail which is the parent of Email Status.
- Similarly, Employee Sponsored Mail **can affect** Onsite Event due to the presence of the active trail between them.
- Also, if Participation Bucket is evidenced, then, The V-Structure will be activated and **influence** can flow between Payroll Deduct and Employee Sponsored Mail.

Agenda

- What are Probabilistic Graphical Models (PGM)
- **PGMs on Sales Ratio Data**
- PGM with respect to CART

Our work in 3 modules:

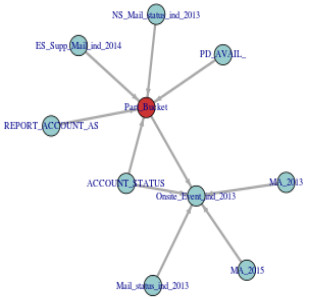
Module 1:



Module 2: Pivot Approach Vs. Bayesian Network Approach

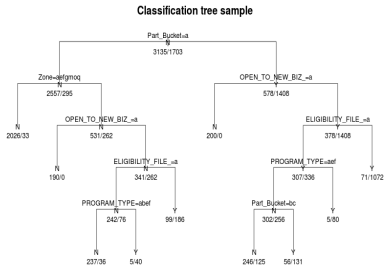
Sno	No. of M	Variable Name	levels				MAP Query Results for Markov E
		par_conv		participant_band	part_perc_bin	MA_2014	Tenure_Years
1	7		low	Low	1	Y	
			medium	High	3	Y	5+
			high	High	3	Y	5+
			extra_high	High	4	Y	5+
		participant_band		ACCOUNT_STATUS	OPEN_TO_NEW_BIZ_	PD_AVAIL_	REPORT_ACCOUNT_AS
2	13		High	ACTIVE	Y	Y	E
			Low	ACTIVE	Y	N	P
		part_perc_bin		PROGRAM_TYPE	ACCOUNT_STATUS	OPEN_TO_NEW_BIZ_	PD_AVAIL_
3	16		1				

Pivots

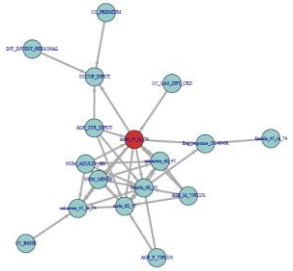


Bayesian network

Module 3: Observations on CART and Bayesina Networks



CART

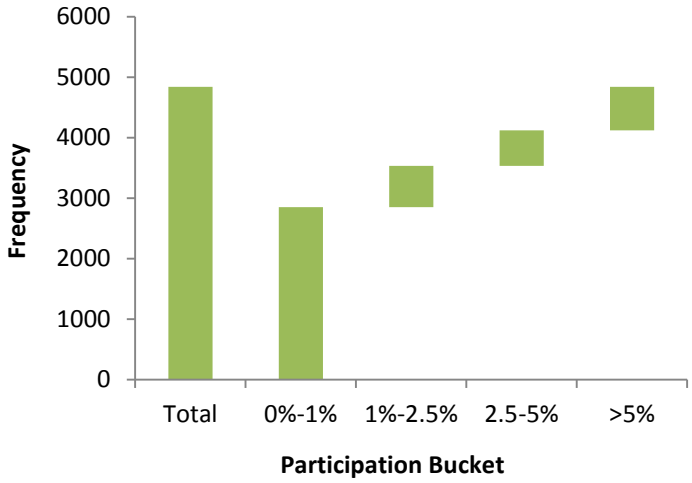


Bayesian Network

Data Details

1. Sales Account Data (2013 to 2015)
2. Demographic details
3. MetLife's mailing
4. Nature and tenure policy
5. Agent and employer type details
6. No. of participants and eligible accounts

1. Target Variable : Participation Bucket
   (Ratio of no. of participants and no. of eligible)
2. No. of observations : 5007



Account Variables
- Company Name
- Market Segment
- Company State

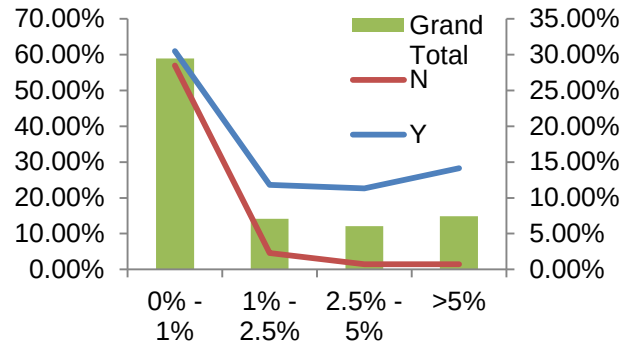
Mail Variables
- Marketing Activity
- Non-Sponsored Mail
- Employee Sponsored Mail

Flag Variables
- New or Existing account flag

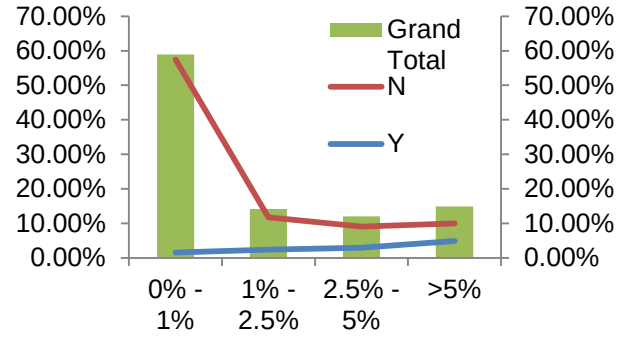
Policy Variables
- Program Type
- Payroll Deduct
- Employer Type

1. Company Name and Company State depict the account details. Market segment could be local, national, regional, USD or other. Majority of the accounts are regional
2. Mailing variable(Non Sponsored Mail) is crucial for participation bucket
3. Policy Variable (Payroll Deduct) is essential to improve participation bucket
4. Most of the accounts are of Single Carrier type and don't have Payroll Deduct

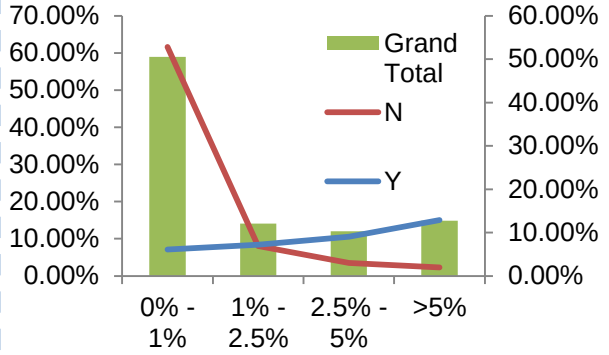
Bivariate Analysis on Target Variable "*Participation Bucket*"



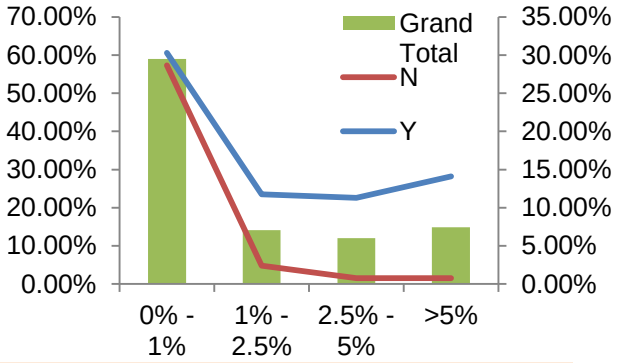
Marketing Activity 2013 - For Participation Bucket in range 2.5%-5%, 11% accounts have marketing activity in 2013



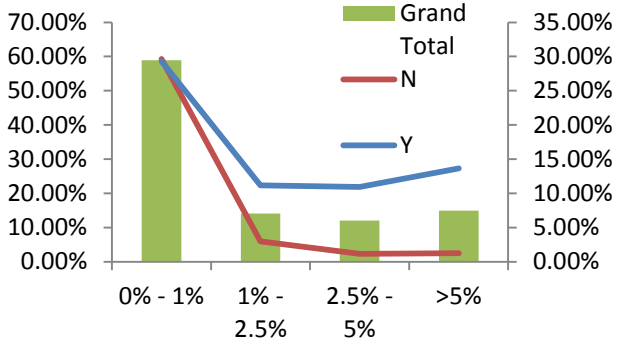
Onsite Event 2013 - For Participation Bucket in range 2.5%-5%, 3% accounts have Onsite Events in 2013



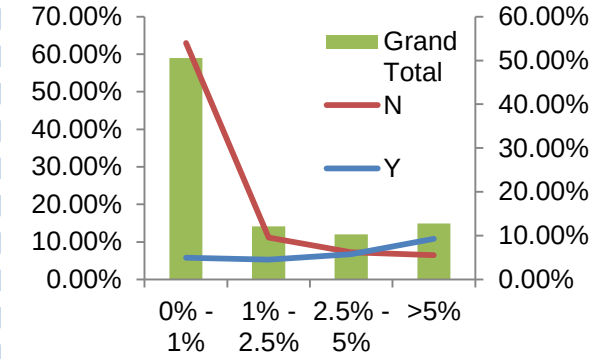
Payroll Deduct - For Participation Bucket in range 2.5%-5%, 9% accounts have Payroll Deduct



Mail Status 2013 - For Participation Bucket in range 2.5%-5%, 11% accounts have Mail Status 2013

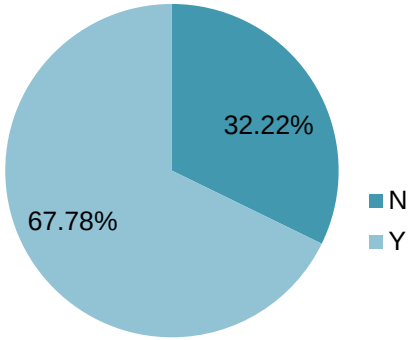


Non Sponsored Mail 2013 - For Participation Bucket in range 2.5%-5%, 11% accounts have Non Sponsored Mail 2013



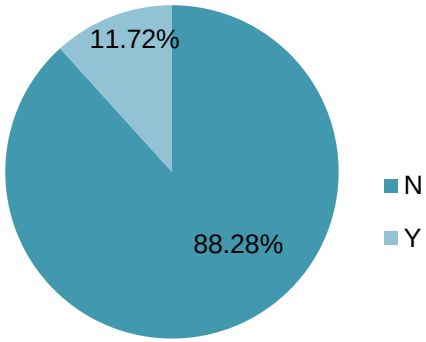
Employee Sponsored Mail 2013 - For Participation Bucket in range 2.5%-5%, 5.81% accounts have Employee Sponsored Mail 2013

Profiling of Important Variables against Target Variable – "*Participation Bucket*"



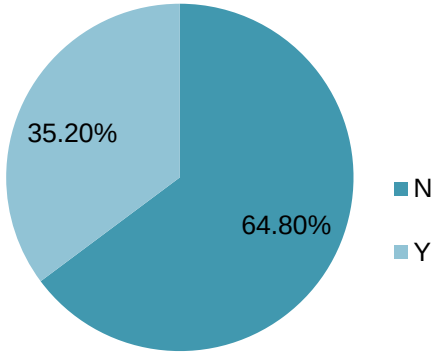
Marketing Activity 2013

68% of the accounts have Marketing Activity 2013



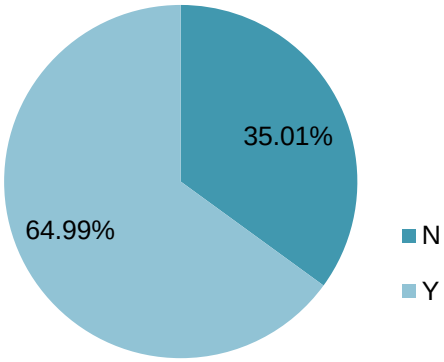
Onsite Event 2013

88% of the accounts don't have Onsite Event 2013



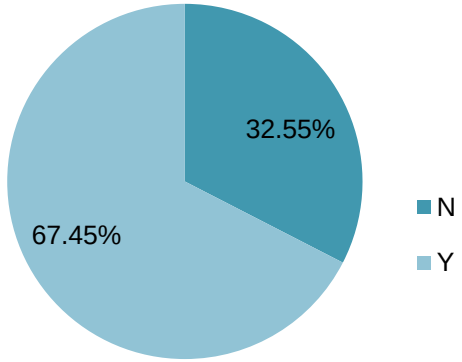
Payroll Deduct

65% of the accounts don't have Payroll Deduct



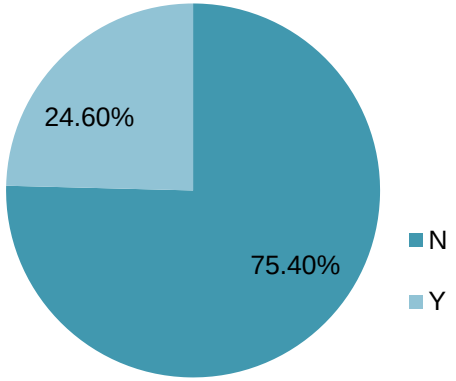
Non Sponsored Mail 2013

65% of the accounts have Non Sponsored Mail 2013



Mail Status 2013

65% of the accounts have Non Sponsored Mail 2013



Employee Sponsored Mail 2013

25% of the accounts received employee sponsored mails.

Bayesian Network discovers NOT ONLY the variables identified by the business users AND ALSO discovers other most influencing variables

Description: To show how to analyze the variables in a Bayesian Network, we tried **to find the values of the variables that influence the target variable the most**, so as to maximize/ minimize the desired target variable class.

The following are the **variables found using the Markov Blanket** of the Target Variable. When considered together, these make the target Variable independent of the rest of the network. Thus, it can be said that **these hold the maximum influence over the target**.

Target Variable:
Participation Bucket

Level 1

Business Users' variables

1st Level Markov Blanket Variables:

- **Payroll Deduct**
- **Non Sponsored Mail 2013**
- **Marketing Activity 2013, 2015**
- Mail Status 2013
- Employee Sponsored Mail 2014
- Account Type
- Account Status
- Onsite Event 2013

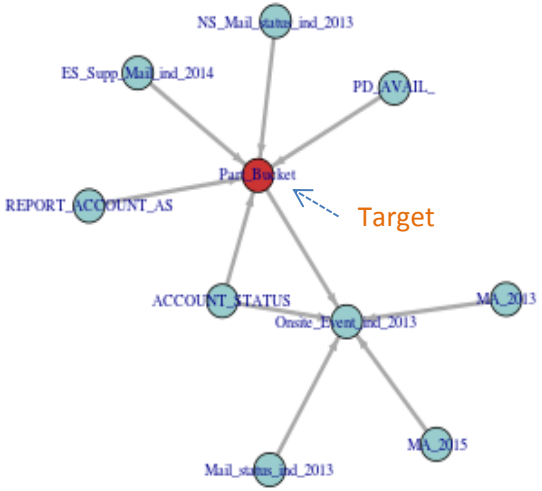
Level 2

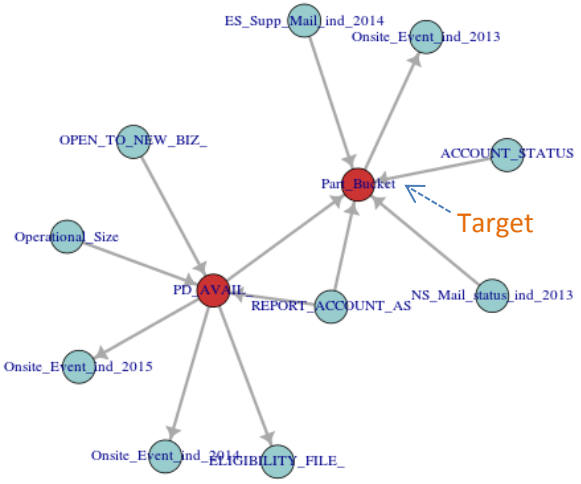
2nd Level Markov Blanket Variables:

- Eligibility File
- **Non Sponsored Mail 2013**
- Employee Sponsored Mail 2014
- **Program Type**
- **Tenure Years**

Business Users' variables

Business Users' variables

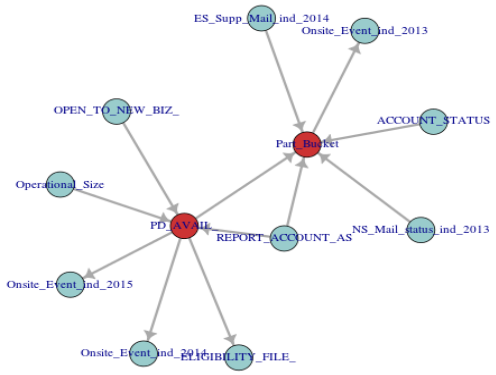




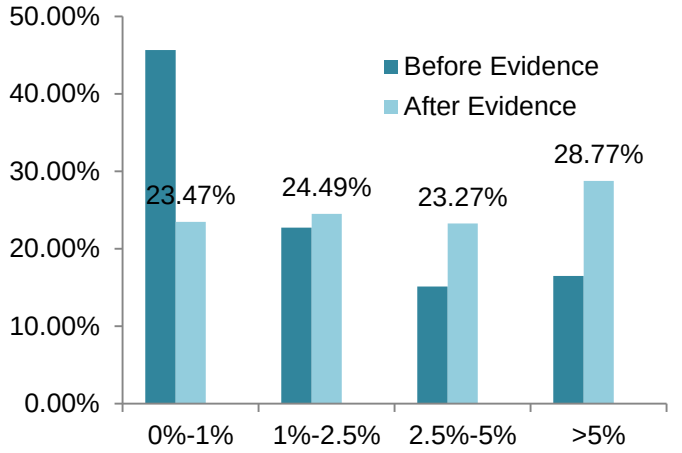
Better Results with Bayesian Network (All Variables)

Description: To show how to analyze the variables in a Bayesian Network, we tried **to find the values of the variables that influence the target variable the most**, so as to maximize/ minimize the desired target variable class.

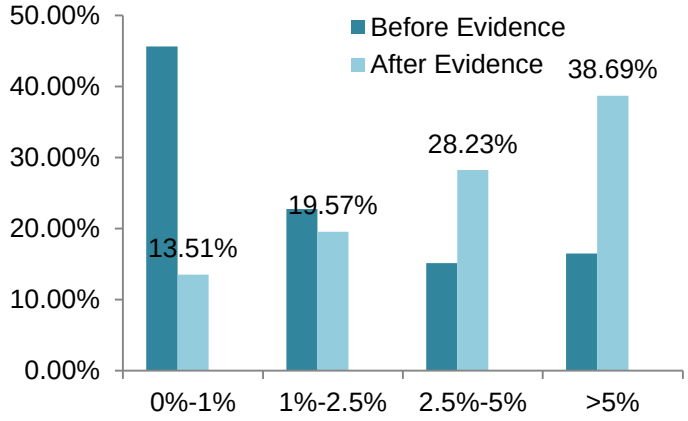
The following are the **variables found using the Markov Blanket** of the Target Variable. When considered together, these make the target Variable independent of the rest of the network. Thus, it can be said that **these hold the maximum influence over the target**.



Variables Selected: Evidence
Payroll Deduct -> Yes



Variables Selected: Evidence
Payroll Deduct -> Yes
Non- Sponsored Mail Status 2013 -> Yes
Mail Status 2013 -> Yes



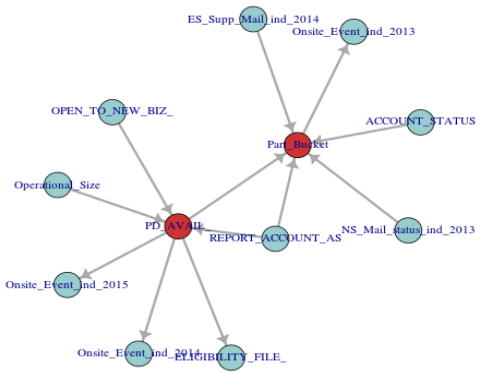
Inference:

- Payroll Deduct is Critical to account performance
- Payroll Deduct without the support of any Mailing Activity cannot drive account performance.

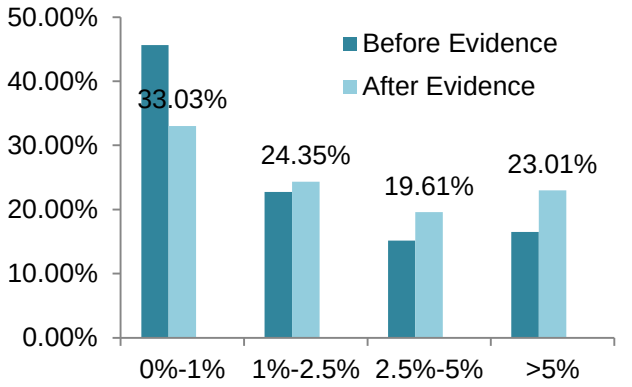
Full Model Analysis on Bayesian Network

Description: To show how to analyze the variables in a Bayesian Network, we tried **to find the values of the variables that influence the target variable the most**, so as to maximize/ minimize the desired target variable class.

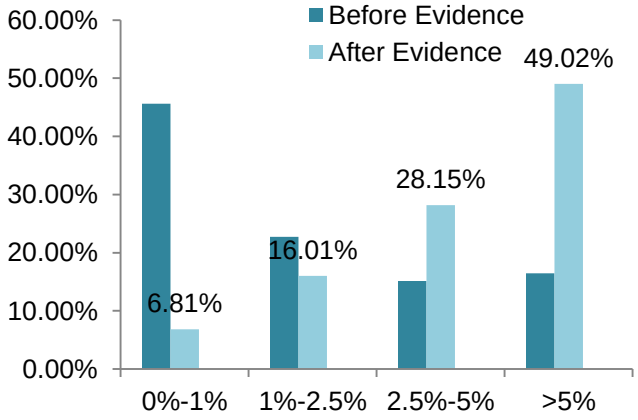
The following are the **variables found using the Markov Blanket** of the Target Variable. When considered together, these make the target Variable independent of the rest of the network. Thus, it can be said that **these hold the maximum influence over the target**.



Variables Selected: Evidence
Eligibility File -> Yes



Variables Selected: Evidence
Eligibility File -> Yes
Non- Sponsored Mail -> Yes
Employee Sponsored Mail ->Yes



Inference:

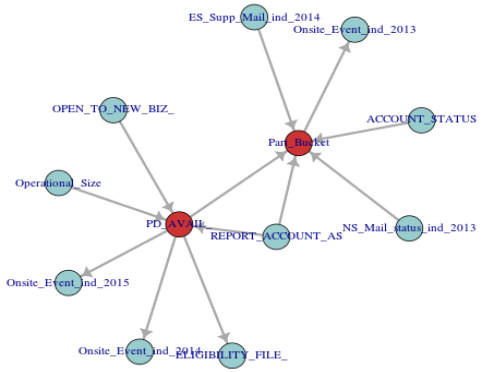
- Payroll Deduct causes the maximum change in the Target Variable. Apart from that, Payroll Deduct's Markov blanket variables like Eligibility File, Employee Sponsored Mail should be considered as well.

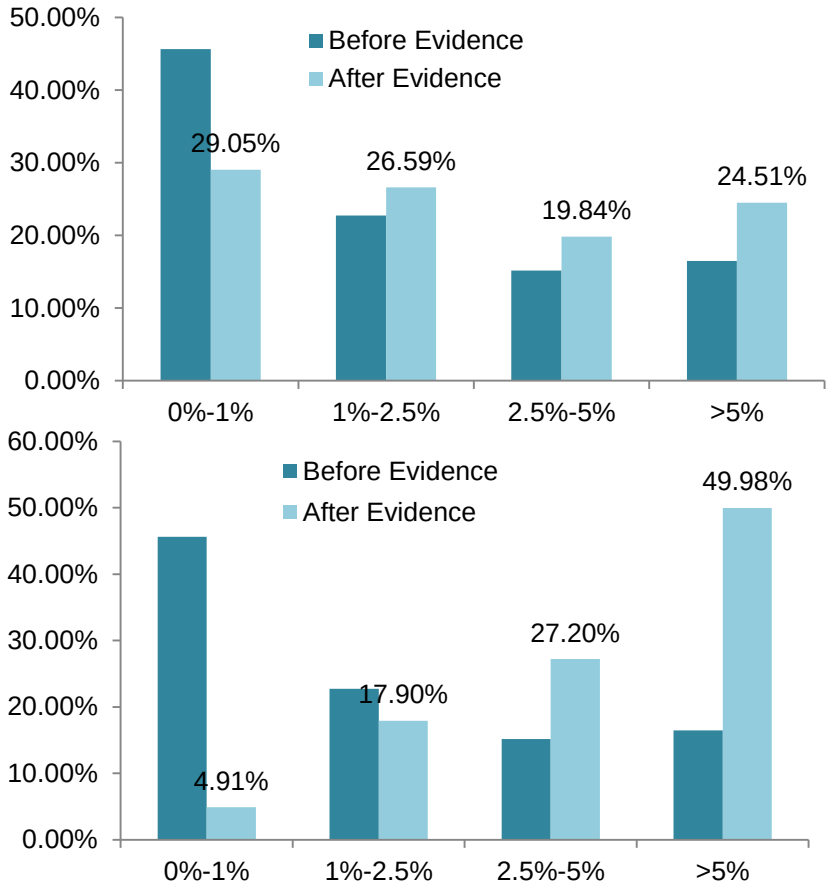
Full Model Analysis on Bayesian Network

Hypothesis: **Mailing activity** significantly **improves participation bucket** with mailing activity in 2013 having more influence than 2014

Variables Selected: Evidence
Employee Sponsored Mail 2014 -> Yes
Non-Sponsored Mail 2014 -> Yes
Onsite Event 2014 -> Yes

Variables Selected: Evidence
Employee Sponsored Mail 2013 -> Yes
Non-Sponsored Mail 2013 -> Yes
Onsite Event 2013 -> Yes





Inference :

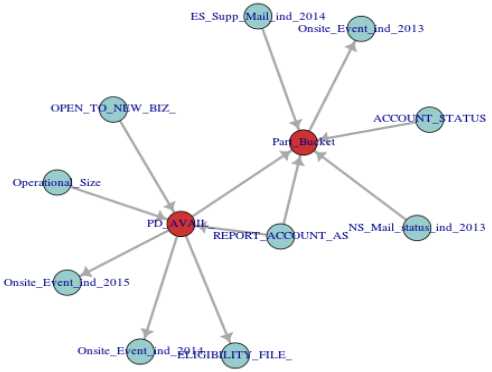
- Improving Mailing Activity for the year 2013 has more influence than in the year 2014
- Certain combinations of marketing activities tend to more effective, in particular - a combination of Non-sponsored, employee sponsored and On-site Events for the year 2013 is yielding higher Participation Bucket (12.05% higher)

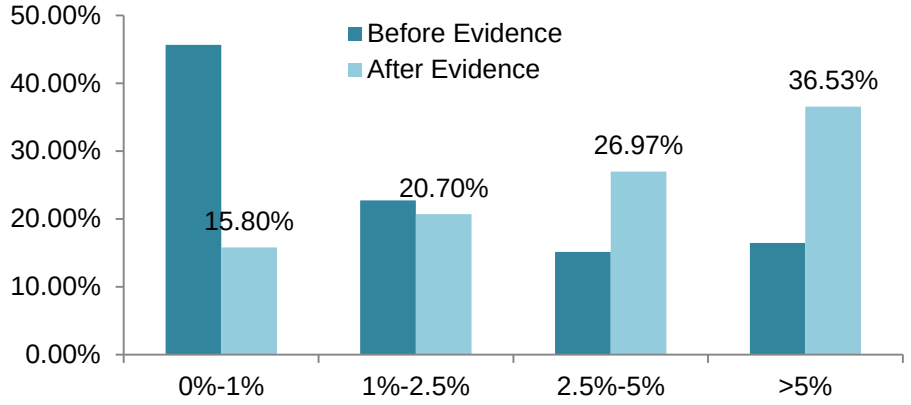
Full Model Analysis on Bayesian Network

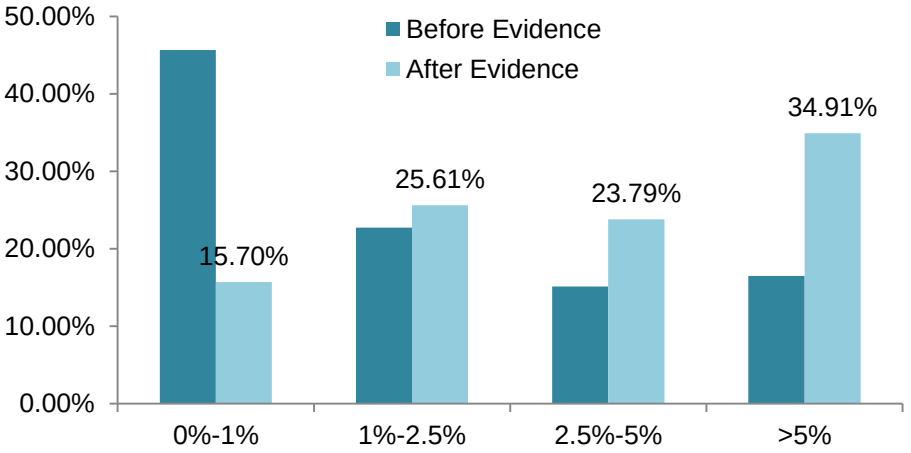
Hypothesis: Participation Bucket **depends on type of Marketing Channel** – Eligibility File, Employee Sponsored Mail, Onsite Events, Payroll Deduct.

Variables Selected: Evidence
Marketing Activity 2013 -> Yes
Payroll Deduct -> Yes

Variables Selected: Evidence
Marketing Activity 2013 -> Yes
Onsite Event 2013 -> Yes





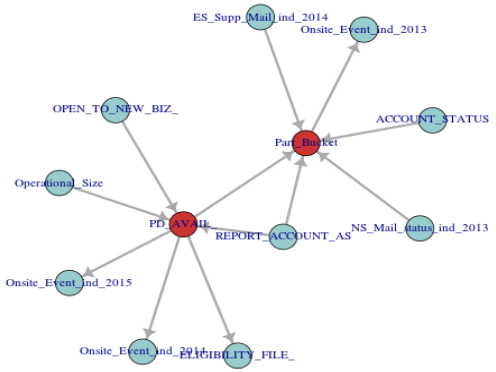


Inference :

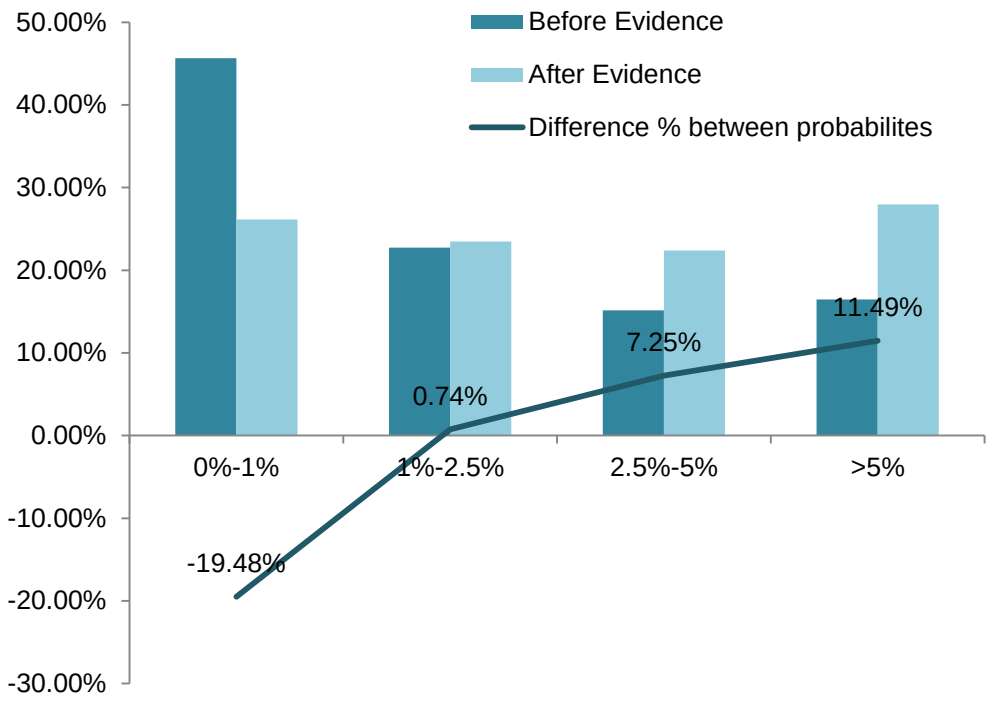
- Along with Marketing Activity, performance also depends on the type of channels – Payroll Deduct, and Onsite Event Activity 2013 being the most important

Full Model Analysis on Bayesian Network

Hypothesis: Newer accounts can improve their Participation Bucket by Mailing Activity



Variables Selected: Evidence
Tenure Years -> 0
Non-Sponsored Mail 2013 -> Y
Employee Sponsored Mail 2013 -> Yes
Onsite Event 2013 -> Yes



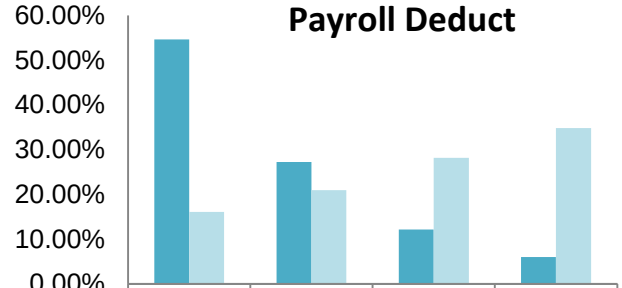
Inference :

- Newer accounts can improve their Participation Bucket by Mailing Activity - specifically, sending non sponsored and employee sponsored mails and doing some onsite events.

Same insights from the both Pivot and Bayesian Analysis (3 variables)

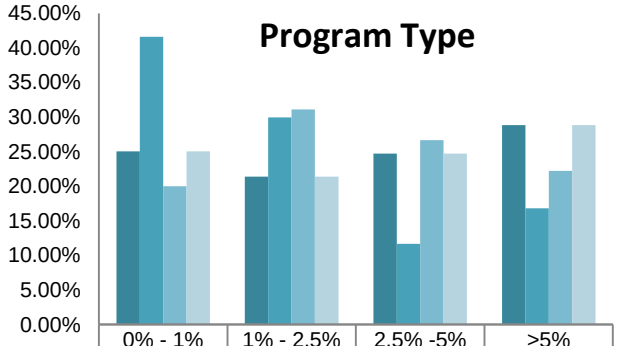
Description: To **test the accuracy** of the Bayesian network, we built models with only those variables that were considered in pivot analysis. We then **observed the target variable** i.e. the Participation Bucket **with respect to selected variables**. The following results were obtained:

Pivot Analysis



	0% - 1%	1% - 2.5%	2.5% - 5%	>5%
Pivot N	54.64%	27.22%	12.16%	5.98%
Pivot Y	16.09%	20.96%	28.16%	34.79%

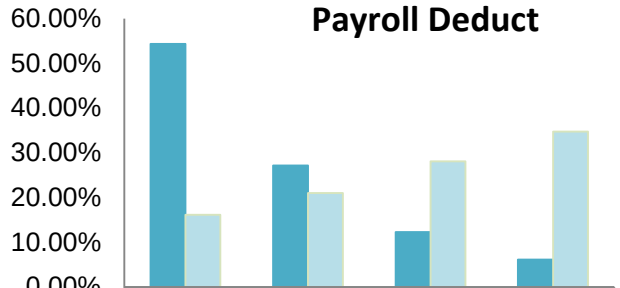
Participation Bucket



	0% - 1%	1% - 2.5%	2.5% -5%	>5%
BROKER CHOICE	25.04%	21.41%	24.74%	28.81%
MARSH COMBINED	41.61%	29.93%	11.68%	16.79%
METLIFE CHOICE	20.00%	31.11%	26.67%	22.22%
BROKER CHOICE	25.04%	21.41%	24.74%	28.81%

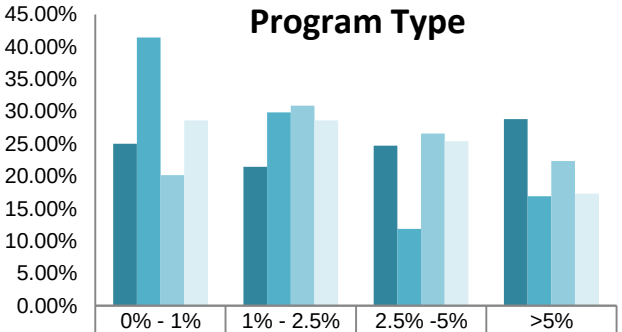
Participation Bucket

Bayesian Analysis



	0% - 1%	1% - 2.5%	2.5% - 5%	>5%
Bayesian N	54.34%	27.19%	12.30%	6.17%
Bayesian Y	16.16%	21.01%	28.11%	34.72%

Participation Bucket



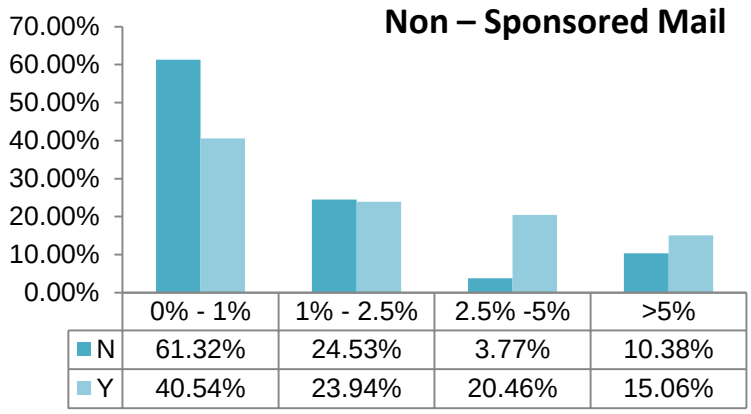
	0% - 1%	1% - 2.5%	2.5% -5%	>5%
SINGLE CARRIER	25.02%	21.46%	24.72%	28.80%
MARSH COMBINED	41.41%	29.87%	11.84%	16.88%
METLIFE CHOICE	20.18%	30.89%	26.61%	22.32%
BROKER CHOICE	28.64%	28.64%	25.40%	17.32%

Participaton Bucket

More insights with more interactions: Pivot and Bayesian (6 variables)

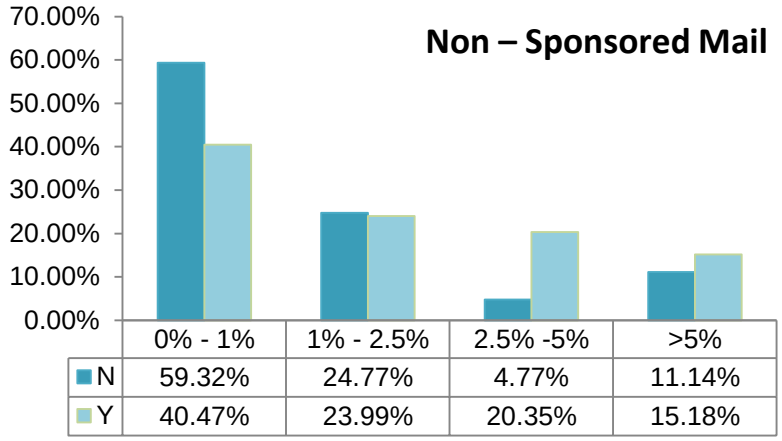
Description: To **test the accuracy** of the Bayesian network, we built models with only those variables that were considered in pivot analysis. We then **observed the target variable** i.e. the Participation Bucket **with respect to selected variables**. The following results were obtained:

Pivot Analysis



	0% - 1%	1% - 2.5%	2.5% -5%	>5%
N	61.32%	24.53%	3.77%	10.38%
Y	40.54%	23.94%	20.46%	15.06%

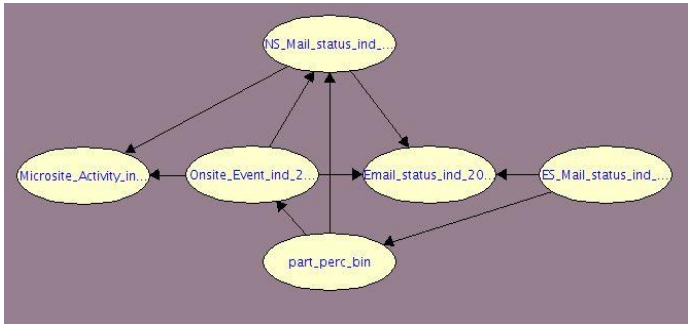
Bayesian Analysis



	0% - 1%	1% - 2.5%	2.5% -5%	>5%
N	59.32%	24.77%	4.77%	11.14%
Y	40.47%	23.99%	20.35%	15.18%

Variables Selected:

- Non Sponsored Mail 2014
- Employee Sponsored Mail 2014
- Onsite Event 2014
- Microsite Activity 2014
- Participation Bucket 2014
- Email Status 2013



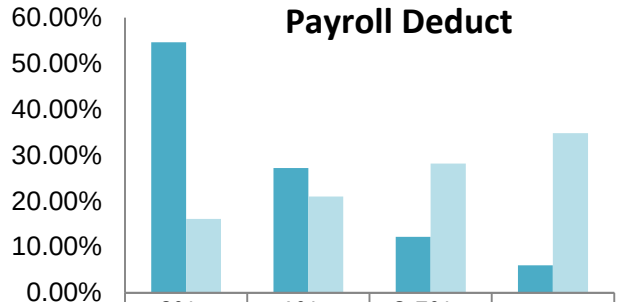
Bayesian Advantage:

- Pivot Tables often return zero values in case it cannot return values according to the set filter when multiple filters are applied.
- Bayesian Networks do not work on values but on conditional probabilities. Thus, even though data rows may be unavailable, probability of the occurrence of each bucket according to the evidence given to filter variables in the network will still be returned.

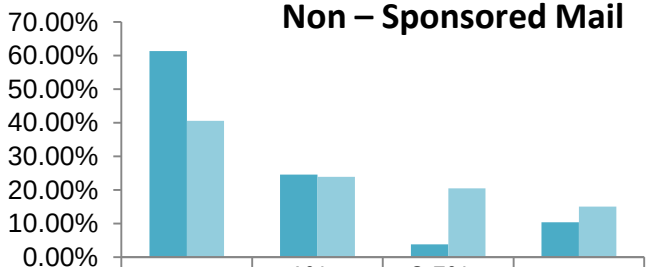
Better Insights with Bayesian (Interaction among all variables)

Description: To **test the accuracy** of the Bayesian network, we built models with only those variables that were considered in pivot analysis. We then **observed the target variable** i.e. the Participation Bucket **with respect to selected variables**. The following results were obtained:

Pivot Analysis

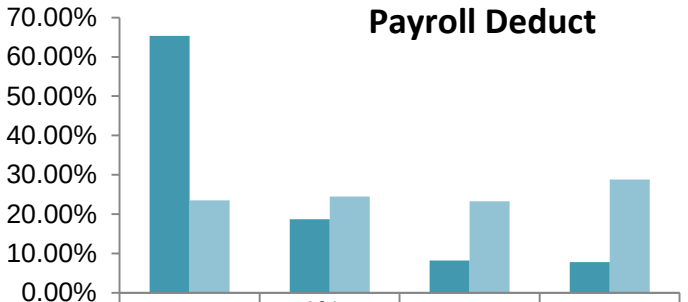


	0% - 1%	1% - 2.5%	2.5% - 5%	>5%
Pivot N	54.64%	27.22%	12.16%	5.98%
Pivot Y	16.09%	20.96%	28.16%	34.79%

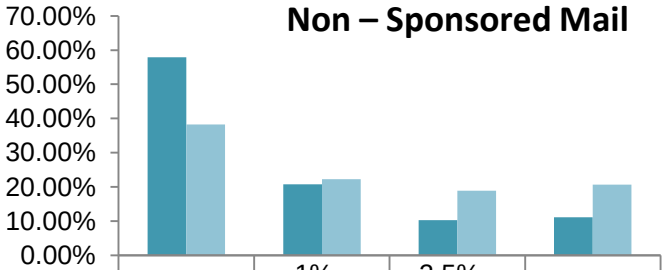


	0% - 1%	1% - 2.5%	2.5% - 5%	>5%
N	61.32%	24.53%	3.77%	10.38%
Y	40.54%	23.94%	20.46%	15.06%

Bayesian Analysis



	0% - 1%	1% - 2.5%	2.5% -5%	>5%
N	65.33%	18.70%	8.19%	7.78%
Y	23.47%	24.49%	23.27%	28.77%



	0% - 1%	1% - 2.5%	2.5% - 5%	>5%
N	57.85%	20.74%	10.30%	11.11%
Y	38.28%	22.24%	18.85%	20.62%

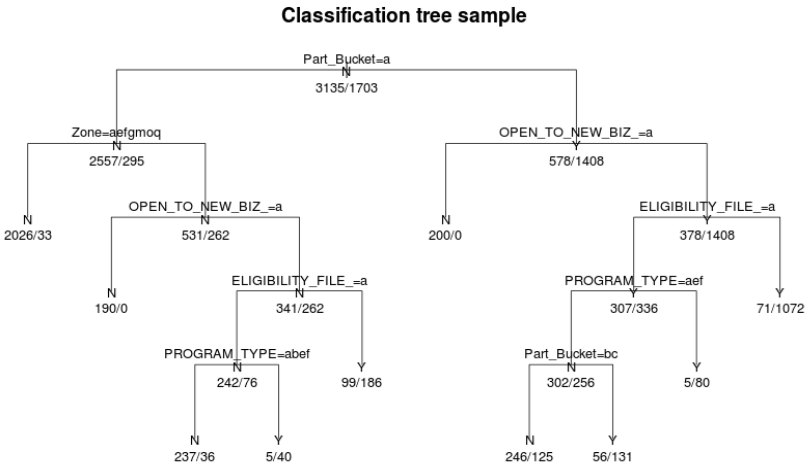
Agenda

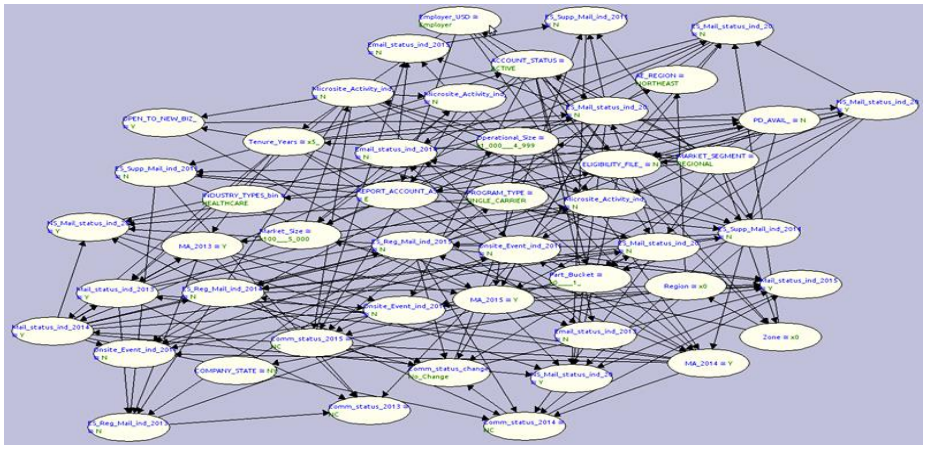
- What are Probabilistic Graphical Models (PGM)
- PGMs on Sales Ratio Data
- **PGM with respect to CART**

Observations On Bayesian Networks and CART

CART is target variable dependent. Problem specific and rigid in terms of modification.

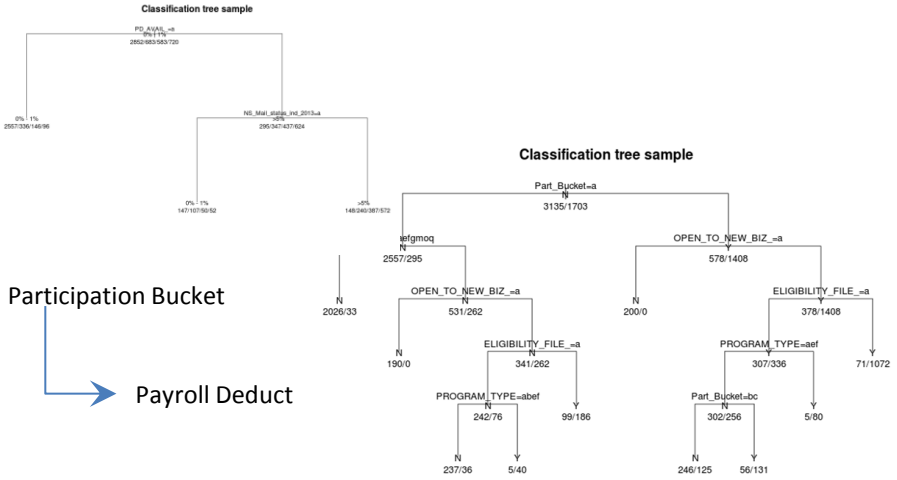
Bayesian **Network doesn't involve fixed target variable**. It is **easier to build and modify**.

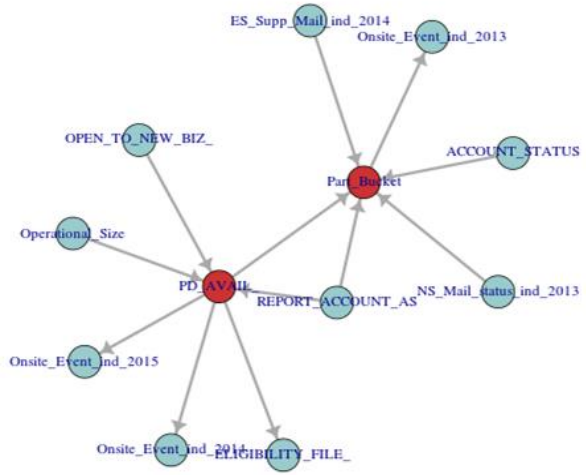




To perform second level analysis we need to construct a new classification tree taking important variables as a target variable.

Second Level of analysis can be done by considering Markov Blanket of the important variable in target's Markov Blanket.



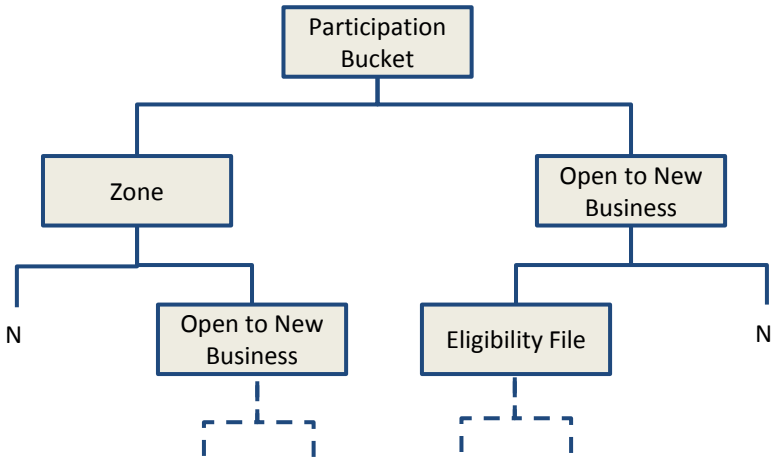


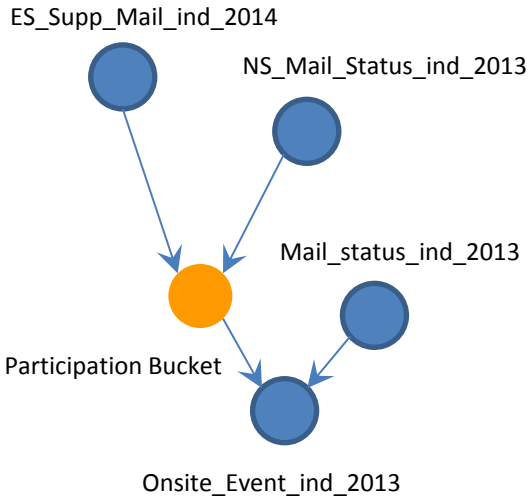
No variable can be added to the tree forcibly or otherwise if needed.

As all variables of data set are already in the model, such need would never arise.

Observations On Bayesian Networks and CART

CART (Classification And Regression Tree): For the same data set, we try to see how many business questions can the Bayesian Network and CART models answer.





Questions	CART	Bayesian Network
If we know that our participation is high, then find responsible variables and their values to replicate the result in the future.	Yes	Yes
For my participation bucket to lie in the higher bucket, should I offer payroll deduct or not?	Yes	Yes
I have not done any mailing activity this year. How did this affect my participation percentage?	No	Yes
To check for conditional dependencies like mailing activity not being helpful if the account is already active.	No	Yes
Is the model easily interpretable?	Yes	Yes

Further Steps

- Optimization of Bayesian Network
- CART comparison