Paper 1169-2017

Using Graph Analytics for Predictive Modeling in Life Insurance

Robert Moore, Thrivent Financial, Minneapolis, MN

ABSTRACT

This paper discusses a specific example of utilizing graph analytics or social network analysis (SNA) in predictive modeling in the life insurance industry. The methods of social network analysis are applied to agents that share compensation, and the results are used to derive input variables for a model to predict the likelihood of certain behavior by insurance agents. Both SAS® code and SAS® Enterprise Miner™ are used to illustrate implementing different graph analytical methods. This paper assumes the reader is familiar with the basic process of creating predictive models using multiple (linear or logistic) regression, and, in some sections, familiarity with SAS Enterprise Miner.

INTRODUCTION

Life insurance and other financial services companies that use field agents to sell and support their products often create predictive models to help manage the agent field force. Such models include retention models to identify agents likely to leave the company in the near future, productivity models to predict the level of agent sales over some future time period, models to identify agents that are most likely to not adhere to required business practices, and in the most serious case models to predict fraudulent behavior. Effective use of such models benefits customers, agents, and the financial services companies by increasing customer satisfaction, improving agent efficiency, and protecting company reputations.

Graph analytic (or social network analysis) methods can be used to derive variables for such predictive models used in the life insurance industry. The example discussed in this paper is a model to predict agent behavior of one of the types mentioned above, where the response (dependent) variable is a binary (0 or 1) outcome. The input (independent) variables for the model included agent specific attributes like their tenure, professional designations, past sales production, and attributes about their customer base. The attributes about their customer base includes things like the number of customers, average customer age, average customer tenure, etc. These variables are fairly typical for use in agent level models. Additional less typical variables were derived from applying graph analytic (social network analysis) methods that rely on connections between the agents. For this model the connections between the agents was shared compensation. Both SAS code and SAS Enterprise Miner were used to derive the graph analytic related variables. The variable names and values displayed in this paper were altered to protect company information.

SOCIAL NETWORKS BASED ON SHARED AGENT COMPENSATION

Life insurance agents often work together to combine expertise in areas in which they specialize. For example, one agent who is not expert in using life insurance for estate planning might require the expertise of another agent who is an expert in using life insurance for estate planning. The agents typically share the selling commission between them in some appropriate proportions like 50%-50% or 70%-30%. The shared compensation records then provide the data to identify the relationship between the agents, and that data can be analyzed using graph analytic techniques.

BASICS OF GRAPH THEORY FOR SOCIAL NETWORK ANALYSIS

Graph analytics for social network analysis are based on mathematical graph theory, and only a very brief non-mathematical introduction to some of the basic terms and definitions from the subject are given here. A mathematical graph can be considered a collection of points connected by lines between some or all of the points. The points and lines together are considered the graph. Other terminology used for the points connected by lines include: nodes connected by links, or vertices connected by edges. For most social network applications, the points represent people and the lines connecting them represent some social relationship (like being friends, or calling each other, etc.). For this application, the points represent agents, and the lines connecting two agents represent shared compensation.

When the lines between the points represent a relationship that has a direction (like employees reporting to a manager) then the graph is called a directed graph. When the relationship has no direction, then the graph is called

undirected. For this application, the lines connecting two agents represent shared compensation which is assumed to have no direction.

The lines between points can be assigned weights to represent the importance or value of the connections between the points. When the lines between the points are weighted, the graph is called a weighted graph. For example, in this application, the lines between two agents could be assigned weights according to how many times they shared compensation or weights indicating how much total compensation the two agents shared.

SHARED COMPENSATION NETWORKS

For this application, the insurance agents are represented by points (nodes) and the lines (links) between agents represent shared compensation. Using the terms defined above, the resulting graph is an undirected, weighted graph that represents a shared compensation network. If a meaningful percentage of agents share compensation, then analysis of the resulting shared compensation network can provide useful input to predictive models.

Insurance companies may have thousands of agents, and the full shared compensation network is actually made up of many sub-networks formed by a limited number of agents that tend to work in the same geographic area. Most of the shared compensation sub-networks are small networks with less than 10 agents, and many consist of just a few agents. Agents that did not share any compensation would be isolated points and are not shown. Figure 1 shows an example of some sub-networks, which was generated using the link analysis node in SAS Enterprise Miner.

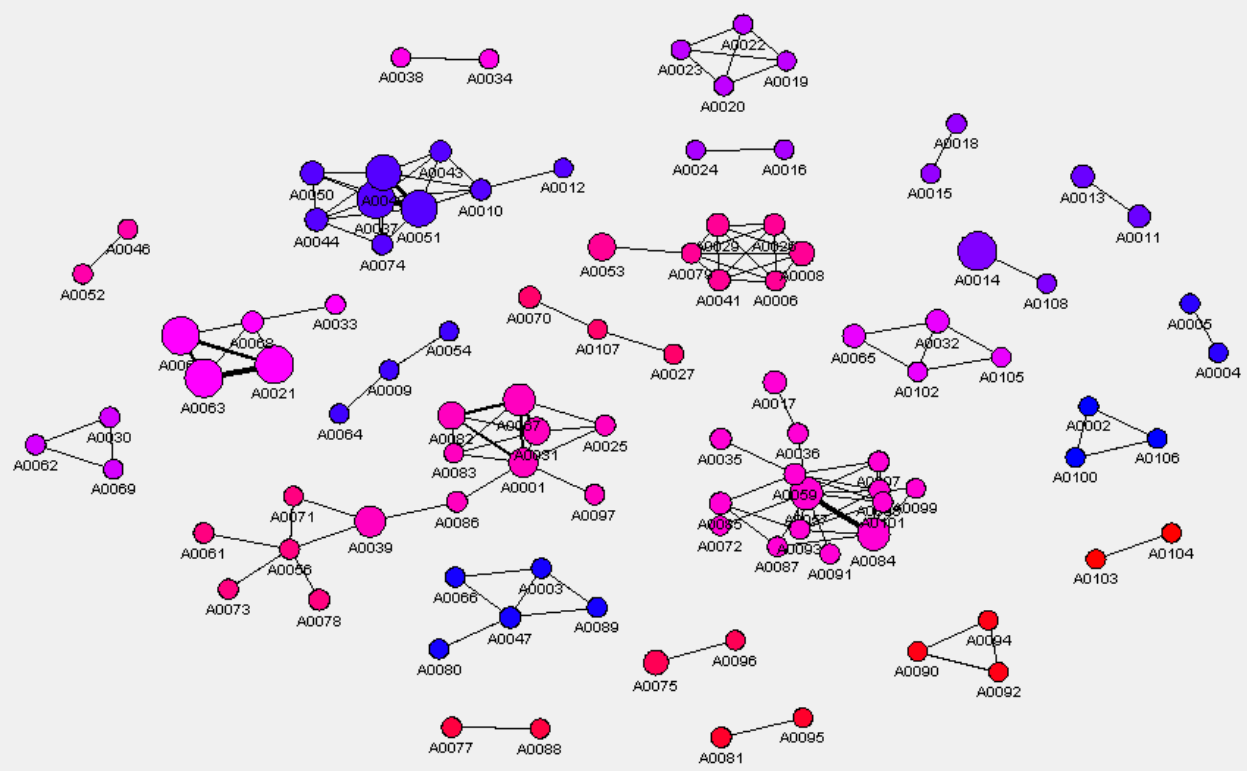

Figure 1 – A selection of agent shared compensation sub-networks.

DATA PREPARATION CONSIDERATIONS

When two agents share compensation on a policy sold to a customer, there are usually two compensation records generated for the policy. So at a minimum this provides the same contract number on the two records and the two agent numbers on the respective records. Figure 2 shows an example where 3 agents shared compensation on sales to 3 different customers. The table on the left shows the summarized compensation records, and the graph on the right shows the corresponding compensation network. The circles represent the 3 agents, and the lines connecting the agents are weighted by the number of times the agents shared compensation. For example, agent 1 and agent 2 shared compensation 2 times, so the line between their points has a weight of 2, whereas agent 2 and agent 3 shared compensation only 1 time, so the line between their points has a weight of 1.

Record	CustID	AgentID
1	CID00001	A0001
2	CID00001	A0002
3	CID00002	A0001
4	CID00002	A0003
5	CID00003	A0001
6	CID00003	A0002
7	CID00003	A0003

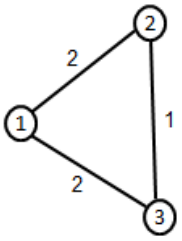

Figure 2 – Example of shared compensation data and corresponding agent shared compensation network.

The timeframe during which the agents shared compensation may affect the number of connections and the weights. Typically, the longer the timeframe considered, the more compensation is shared, so there may be more agents that shared compensation, more connections between agents, and higher weights on the connections. The timeframe for the predictive model discussed in this paper used the 6-month period prior to the response observation window.

Another consideration is whether to include only active agents or both active and terminated agents. Some agents may have left the company before the end of the timeframe, but may have shared compensation with other agents prior to leaving. This may be important if shared compensation data is used in agent attrition models, or models to predict agent productivity.

What values to use for the weights on the connections is also an important consideration. It may provide insight to try different weighting schemes. For example, simple counts of how many times two agents shared compensation versus the total amount of compensation shared.

SOCIAL NETWORK ANALYSIS METRICS AS INPUT VARIABLES

The agent shared compensation network was analyzed using the SAS Enterprise Miner link analysis node, and the resulting summary social network metrics produced by the node were used as input variables for each agent. Additional related metrics were also derived using SAS code and summarized as input variables for the predictive model.

ENTERPRISE MINER LINK ANALYSIS NODE

The link analysis node in SAS Enterprise Miner was used to produce social network analysis metrics. The link analysis node presents the shared compensation network in visual form as a mathematical graph with the points (nodes) representing agents and the lines (links) connecting the points indicating shared compensation. Because thousands of agents were considered in this example, the entire shared compensation network was too dense to see anything meaningful, but looking at subsets of the graph reveals more detail. By focusing on specific geographic areas, the numerous sub-networks are clearly seen as shown in figure 1.

The structure of the dataset used within the link analysis node was similar to the example shown in figure 2. Within Enterprise Miner the role of the dataset was set to ‘transaction’, and the role of the customer ID was set to ‘ID’, and the role of the agent ID was set to ‘target’ as shown in figure 3.

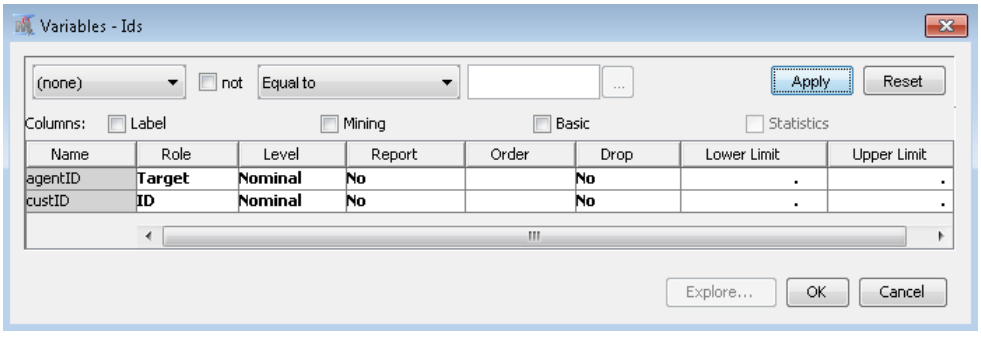

Figure 3 – The variables roles in the link analysis node in SAS Enterprise Miner.

The lower right portion of figure 4 shows a portion of the diagram window with the link analysis node connected to the dataset. The lower left corner shows the properties panel for the link analysis node. In the properties panel, the minimum confidence percentage is shown set to 1%, and the association support type is set to 'Count', and the association support count is set to 1 to include as many shared compensation associations as possible. It is useful to experiment with these settings as they will determine how many connections will be included in the network.

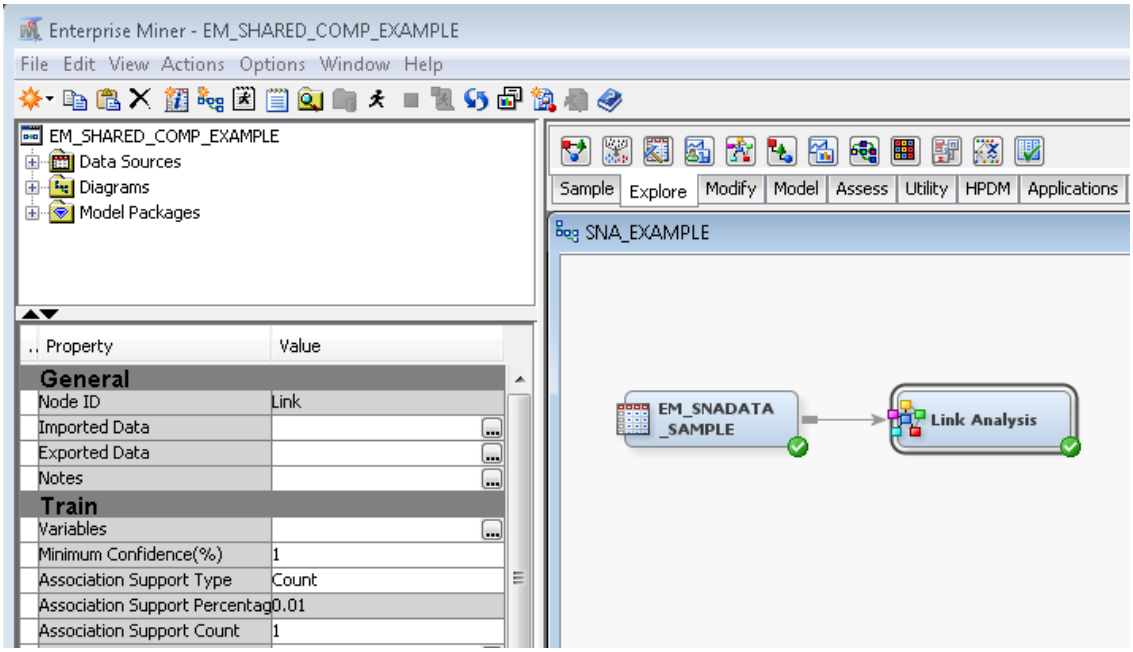

Figure 4 – The input dataset and link analysis node in SAS Enterprise Miner.

Running the link analysis node provides visual representations of the network and tables with a variety of SNA metrics for the network. For example, the sub-networks shown in figure 1 are from the constellation plot produced by the link analysis node. Figure 5 shows some of the SNA metrics calculated for the network that can be found in the node data behind the item constellation plot. This data can be seen by selecting the "Item Constellation Plot" window in the output, and then selecting View, and then selecting Table. The data can then be saved as a SAS dataset. In the leftmost column of the table are the nodes which in this example are agent ID numbers, and for each agent SNA centrality metrics are shown. This table was saved and then merged by agent ID with the other input variables.

Node	Weight	Out-degree Centrality	Weighted Eigenvector Centrality	Unweighted Eigenvector Centrality	Weighted Closeness Centrality	Unweighted Closeness Centrality	Weighted Betweenness Centrality	Unweighted Betweenness Centrality	Weighted Influence1 Centrality	Weighted Influence2 Centrality	Unweighted Influence1 Centrality	Unweighted Influence2 Centrality	Clustering Coefficient Centrality	Item-cluster
A0001	170	7	2.27E-14	0	0.257961	0.18232	0.010307	0.009826	0.068052	0.190812	0.001334	0.004575	0.380952	16
A0002	1	2	0	0	0.234412	0.169521	0	0	.0003812	.0007625	.0003812	.0007625	1	1
A0003	10	3	0	0	0.239981	0.172174	0	.0001031	0.002478	0.009531	.0005719	0.001525	0.666667	2
A0004	2	1	0	1.67E-13	0.23285	0.168081	0	0	.0003812	.0003812	.0001906	.0001906	0	3
A0005	3	1	0	1.64E-13	0.23285	0.168081	0	0	.0003812	.0003812	.0001906	.0001906	0	3
A0006	15	5	0	0	0.244088	0.175221	0	0	0.00305	0.054708	.0009531	0.004956	1	18
A0007	23	4	0	0	0.261377	0.185047	0	.0007706	0.002669	0.098742	.0007625	0.005147	0.666667	15
A0008	92	5	1.7E-14	2.7E-14	0.244566	0.175221	0.002268	0	0.020015	0.037743	.0009531	0.004956	1	18
A0009	7	2	0	0	0.234412	0.169521	.0002061	.0002061	.0003812	.0003812	.0003812	.0003812	0	4
A0010	25	5	0	0.686904	0.250144	0.177738	0.001443	0.001443	0.010865	0.402592	.0009531	0.004956	0.6	5

Figure 5 – The SNA metrics from the link analysis node in SAS Enterprise Miner.

OTHER SHARED COMPENSATION NETWORK BASED VARIABLES

Additional input variables were derived from calculating the overlap ratio (see Zheng, 2011) which is equal to 2 times the number of times compensation was shared between two agents divided by the sum of the number of times compensation was shared by each agent. So if s_{ij} denotes the number of times agents i and j shared compensation,

s_i is the number of times agent i shared compensation, and s_j is the number of times agent j shared compensation, then the overlap ratio $r = \frac{2 \cdot s_{ij}}{s_i + s_j}$.

The overlap ratio reflects the number of times compensation was shared between two agents versus the total number of times they shared compensation separately. The value will be small if two agents rarely shared compensation with each other compared to how many times they shared compensation with other agents. It will be large if two agents shared compensation with each other many times compared to how many times they shared compensation with other agents.

The overlap ratio is calculated between all pairs of agents that shared compensation, so for each agent, overlap ratio summary statistics were calculated in order to incorporate the metric as an input variable for modeling. These derived variables included the mean, maximum, and range of the overlap ratios between the agent and all the other agents with whom they shared compensation.

FIRST DEGREE CONNECTION BASED VARIABLES

The first degree connections for an agent are all the other agents in the shared compensation network that can be reached in just one step (all the agents with whom they shared compensation). Additional variables can be derived by summarizing attributes of the 1st degree connections for each agent (see Baesens, 2014). For example, if the target is a binary response taking the values 1 or 0, then the number and percentage of responders can be calculated over the agents that are the 1st degree connections for a particular agent. Similarly, other agent attributes can be summarized over the 1st degree connections for each agent by using appropriate statistics like the minimum, mean, maximum, range, and standard deviation. For example, agent attributes like tenure, prior sales production, number of customers, etc. can be summarized across the agents that are 1st degree connections, and all these summarized variables can then be used as input variables for modeling.

The second degree connections for an agent are the agents in the shared compensation network that can be reached in exactly two steps. Additional variables can also be derived for these 2nd degree connections.

CONCLUSIONS

Shared compensation between insurance agents or financial service company representatives can be interpreted as a mathematical graph with points (nodes) representing agents and the lines (links) between the agents representing shared compensation. Utilizing the resulting agent shared compensation network and variables derived from the SNA metrics can provide additional input variables to predictive models that provide insight into agent behavioral aspects that other variables may not capture. The Enterprise Miner link analysis node provides SNA centrality measures for agents that can be captured and merged with other modeling variables. Overlap ratios can be calculated directly and input variables derived from summary statistics between agents and the other agents with whom they shared compensation. Finally, agent attributes among the first and second degree connections for an agent can be summarized to provide variables that provide additional behavior insights.

REFERENCES

Baesens, B. (2014) “Analytics in a Big Data World”, Hoboken, NJ: John Wiley & Sons

Chartrand, G. (1985), “Introductory Graph Theory”, Mineola, NY: Dover Publication, Inc.

Chartrand, G., and Zhang, P. (2012), “A First Course in Graph Theory”, Mineola, NY: Dover Publication, Inc.

Liu, Y., Lee, T., Zhang, R., Dean, J. (2014), “Link Analysis Using SAS® Enterprise Miner™", Cary, NC: SAS Institute Inc.

Zheng, J. (2011), "Visualizing Healthcare Provider Network using SAS® Tools ", PharmaSUG2011, Paper HS08

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Robert Moore
Thrivent Financial
625 Fourth Avenue South
Minneapolis, MN 55415
Work Phone: 612-844-4036
E-mail: robert.moore@thrivent.com